EXPLORING THE ROLE OF MITOCHONDRIAL DNA QUANTITY AND QUALITY IN CARDIOVASCULAR DISEASE

by
Ryan Joseph Longchamps

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
August 2019

Abstract

Mitochondria represent an essential component of healthy aging, in part due to their critical role in cellular metabolism. Declines in mitochondrial function have been linked to several aging-related diseases, such as cardiovascular disease (CVD). Although many risk factors for CVD are have been fully elucidated, understanding the underlying role of mitochondrial function in CVD etiology may reveal untapped avenues for clinical testing and treatment. We begin to approach this issue by tackling two novel biomarkers of mitochondrial function: mitochondrial DNA (mtDNA) quantity and quality. First, we examine the impact of mtDNA quantity estimation methods on outcomes to establish a new gold standard for this novel phenotype. We then perform the largest genome-wide association study for mtDNA quantity to date and identify putative loci controlling mtDNA quantity. With these loci we discover previously unidentified pathways which may contribute to control of mtDNA quantity and may explain its link to mitochondrial function. In the final part of this work we examine the impact of poor mtDNA quality on mortality and incident CVD. We establish accumulated mtDNA mutations as a risk factor of mortality and CVD, independent of traditional risk factors. Finally, we assess how the mutational burden of mtDNA mutations modulates risk for mortality and CVD.

Advisor: Dan E. Arking, Ph.D.

Reader: Brian O'Rourke, Ph.D.

Acknowledgments

First, I would like to thank my mentor Dan Arking. Dan has been an amazing mentor these past six years, and I know I would not be where I am today, both as a scientist and as a person, were it not for him. Dan seemingly taught me everything I know about computational genetics, but gave me the independence to grow as a responsible scientist. I will value his impact on my life for years to come and look forward to catching up at yearly meetings. Thank you, Dan.

Dans ability to put together a team of people to not only do great science together, but also meld into a great group of friends outside of the lab environment is bar none. From the lab I joined with Anna, Foram, Shannon, and Nate to the lab I am leaving with Rebecca, Christina, Stephanie, Thuy Vy, Vamsee, and Charles – thank you all for your mentorship, advice, thoughts, suggestions, jokes, laughs, happy hours and fun.

I would like to thank the members of my thesis committee: Brian O'Rourke, David Valle, Ingo Ruczinski and Loyal Goff for always providing amazing feedback and mentorship throughout my time in graduate school.

To the Human Genetics program, thank you for taking a chance on me. Thank you for creating such a wonderful place to train, learn and grow. This work would not be possible without David Valle, Sandy Muscelli, and Kirby Smith.

I would also like to thank the MD-GEM program. They gave me the cross disciplinary training so necessary and underappreciated in science as well as provided me with several amazing opportunities to travel and meet amazing scientists from all over the world. Thank you Priya Duggal and Jennifer Deal for creating such an amazing program.

To my classmates, thank you for being the best group of people I could ever have imagined going through the program with. Through all of the courses, problem sets, discussion, presentations, and journal clubs we all did this together. I would especially like to thank Genay and Sarah– I would not have made it through third year without you two.

To all of my friends both in and out of Hopkins, thank you for always being the necessary distraction I needed. Mitch, Stav, Nick, Ricky, Scott, Kat, Eric, Justin, Genay, Sarah, Ben, Joel, Anna, Foram, Shannon, you are all amazing people.

To the Echols family, thank you for being my home away from home. You all have been so welcoming these past few years and it is amazing to know I am officially a part of your family.

Thank you to my family. Danielle, you are an amazing big sister. You have taught me so much about this world and I hope to one day exude even a fraction of the compassion and

drive to make a difference that you do every day. My parents, John and Linda, thank you for never pushing me and letting me find my own path. Thank you for always being there, for your unconditional love, and your friendship. I would not be where I am today were it not for the sacrifices you two have made. All of this work is just as much yours as it is mine.

And finally, to my wife, Charlotte. I am forever grateful for your unwavering love and support. Through all of the stress and late nights you have been there for me and I hope to let you know every day how thankful I am.

Table of Contents

List of Tables

Supplementary Tables

List of Figures

Supplementary Figures

Chapter 1 : Introduction

MITOCHONDRIAL DYSFUNCTION IN DISEASE

The mitochondrion is double membraned organelle involved in several cellular processes such as intracellular signaling, reactive oxygen species (ROS) production, cellular differentiation and apoptosis. However, the primary function of the mitochondrion is known as oxidative phosphorylation whereby the chemical energy required for cellular metabolism is created. Oxidative phosphorylation occurs within the mitochondrial cristae where the major byproducts of the citric acid cycle, NADH and $FADH_2$, are oxidized to generate a proton gradient which is subsequently used to generate ATP. Due to its integral role in energy supply, mitochondrial dysfunction and declines in oxidative capacity have long been hypothesized to underlie critical changes which increase vulnerability to chronic disease[1–3].

Of interest to our lab, mitochondrial dysfunction has specifically been linked to cardiovascular disease (CVD)[4], one of the leading causes of mortality and morbidity in the United States. CVD represents a complex cluster of diseases of the blood vessels and heart that often reveals clinically as myocardial infarction and stroke. In spite of its heterogenous nature, CVD is typically driven by endothelial damage and a subsequent chronic inflammatory response known as atherosclerosis where a buildup of fats and cholesterol creates arterial plaques which can restrict blood flow or burst. Mitochondrial dysfunction has been linked to atherosclerosis by affecting plaque instability and blood clot formation through changes in ROS production, apoptosis and intracellular signaling[5,6]. Additionally,

mitochondrial dysfunction in mice mediated by mitochondrial DNA (mtDNA) damage demonstrated increased a larger necrotic core and thinner fibrous cap, key features of plaque instablility[7]. Furthermore, increased ROS production has been linked to endothelial senescence[8,9] and loss of endothelial integrity[10], possibly highlighting a role where mitochondrial dysfunction contributes to the initiation of atherosclerosis.

Although several pathways through which mitochondrial dysfunction acts on CVD have been explored, several questions have yet to be answered. In this work, we hope to address some of these questions by interrogating two minimally invasive biomarker of mitochondrial function: mtDNA quantity[11] and mtDNA quality[12].

MITOCHONDRIAL DNA QUANTITY

Unlike most organelles, the mitochondrion possesses its own genome, an intron-free, double-stranded, 16.6 kb maternally inherited, circular DNA molecule with 37 genes vital to oxidative phosphorylation. The term mtDNA quantity, also known as mtDNA copy number (mtDNA-CN), is in reference to the fact that a large amount of variation exists in the number of copies of mtDNA present within cells, tissues, and individuals. Several methods exist in which to measure mtDNA-CN, however with the advent of several newer methods there has yet to be a rigorous examination of these methods to establish a new gold standard. Without a comprehensive comparison several researchers may be relying on suboptimal methods resulting in misrepresented associations involving mtDNA-CN. Here, we approach this issue by performing an intrinsic validation to provide the field with

qualitative information on the performance of five common mtDNA-CN estimation methods.

Previous research has demonstrated mtDNA-CN directly correlates with mitochondrial function as measured through energy reserves, oxidative stress, and mitochondrial membrane potential[13]. In fact, previous work from our lab has demonstrated individuals in the lowest 20th percentile of mtDNA-CN were 23% more likely to develop CVD compared to individuals in the highest 20th percentile[14]. Additionally, we were able to improve 10-year CVD risk classification when mtDNA-CN was added to the Pooled Cohort Equation from the 2013 American College of Cardiology/American Heart Association guideline on assessment of CVD risk. Importantly, these findings were independent of traditional risk factors indicating mtDNA-CN may be acting orthogonal to these factors in an independent pathway.

In this work, we take a population genetics approach to exploring how mtDNA-CN is associated with CVD independent of traditional risk factors. Although several Mendelian mitochondrial disorders have identified genes which modulate mtDNA-CN, the comprehensive mechanism through which copy number is regulated is largely unknown[15]. To further identify the genetic factors which control mtDNA-CN, we leverage data from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and the UK Biobank to perform a genome-wide association study (GWAS). With more mtDNA-CN loci identified we identify pathways and biological processes which correlate with mtDNA-CN as a means to better understand the role mtDNA-CN plays in CVD.

MITOCHONDRIAL DNA QUALITY

In addition to mtDNA-CN, mtDNA quality, or an accumulation of mtDNA mutations, has also been shown to be a useful biomarker for mitochondrial function[12]. As a natural byproduct of oxidative phosphorylation, ROS lead to mtDNA damage and declines in mitochondrial respiratory activity[16]. These subsequent declines to mitochondrial respiration could further exacerbate ROS levels resulting in a negative feedback loop of mitochondrial dysfunction. In fact, oxidative damage has been shown to markedly increase with age[17,18] showing the importance of this negative feedback loop. Furthermore, oxidative stress induced mtDNA damage persists longer than nuclear DNA damage highlighting the long term effects mtDNA damage can have on subsequent gene expression[19].

Previous research has focused on exploring the impact of specific mtDNA mutations, or heteroplasmies, on mitochondrial function and disease. For example, m.3243A > G, a mutation within the mitochondrial tRNA leucine, has been widely studied and linked to mortality, dementia, stroke[20], and carotid artery dissection[21]. However, little is known about the overall heteroplasmic burden, or DNA quality, on disease. With the wider availability of whole genome sequence data, we are no longer limited to asking questions about single heteroplasmic sites. Instead, we can assess an individual's risk based on heteroplasmies found across the entire mtDNA genome. As such, in the final part of this work we aim to address the impact of overall heteroplasmic mutational burden on mortality and CVD.

Chapter 2 :
Evaluation of mitochondrial DNA copy number estimation techniques

INTRODUCTION

Mitochondrial dysfunction has long been known to play an important role in the underlying etiology of several aging-related diseases, including cardiovascular disease (CVD), neurodegenerative disorders and cancer[4]. As an easily measurable and accessible proxy for mitochondrial function, mitochondrial DNA copy number (mtDNA-CN) is increasingly used to assess the role of mitochondria in disease. Several population-based studies have shown higher levels of mtDNA-CN to be associated with decreased incidence for CVD and its component parts: coronary artery disease (CAD) and stroke[14,22]; neurodegenerative disorders such as Parkinson's and Alzheimer's[23,24]; as well as several types of cancer including breast, kidney, liver and colorectal[25–27]. Furthermore, mtDNA-CN measured from peripheral blood has consistently been shown to be higher in women, decline with age, and correlate negatively with white blood cell (WBC) count[28–30].

Although the mtDNA-CN field is relatively young, the number of publications has been steadily increasing at an average rate of 12% per year since 2015[31]. However, there has yet to be a rigorous examination of the various methods for measuring this novel phenotype and the factors which may influence its accurate estimation. Without such an examination, studies may be severely underestimating or misrepresenting the relationship of mtDNA-CN with their traits of interest.

Quantitative real-time PCR (qPCR) has been the most widely used method for measuring mtDNA-CN, partly due to its low cost and quick turnaround time. However, recent work has demonstrated the feasibility of accurately measuring mtDNA-CN from preexisting microarray, whole exome sequencing (WES) and whole genome sequencing

(WGS) data[14,29,32]. With these advances, it is important for the field to evaluate these methods in the context of the current gold standard.

In addition to the method for determining mtDNA-CN, it is important to consider the impact of DNA extraction method on mtDNA-CN, particularly due to the small size and circular nature of the mitochondrial genome. Previous research has shown organic solvent extraction is more accurate than silica-based methods at measuring mtDNA-CN, which is unsurprising as column kit parameters are typically optimized for DNA fragments $\geq$ 50 Kb[33]. However, as all DNA extraction methods have bias in the DNA which they target, measuring mtDNA-CN from direct cell lysate may prove to be a more accurate method.

In the present study, we assess the relative performance of various methods for measuring mtDNA-CN and the effects of DNA extraction on mtDNA-CN estimation accuracy. We leverage mtDNA-CN calculated across 4,574 individuals from two prospective cohorts, the Atherosclerosis Risk in Communities study (ARIC) and the Multi-Ethnic Study of Atherosclerosis (MESA). Using mtDNA-CN estimates calculated from qPCR, WES, WGS, and two microarray platforms – the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina HumanExome BeadChip genotyping array – we compare associations for known correlates of mtDNA-CN including age, sex, white blood cell count, the Duffy locus and incident CVD to determine the optimal method for calculating copy number. We additionally determined the reproducibility of mtDNA-CN measurements *in vitro* from three separate DNA extraction methods: silica-based column selection, organic solvent extraction (phenol-chloroform-isoamyl alcohol), and measuring mtDNA-CN from direct cell lysis without performing a traditional DNA extraction. We

hypothesized that mtDNA-CN calculated from WGS data would outperform other estimation methods and mtDNA-CN measured from direct cell lysate would be more accurate than traditional DNA extraction methods.

METHODS

Study populations

The ARIC study recruited 15,792 individuals between 1987 and 1989 aged 45 to 65 years from 4 US communities. DNA for mtDNA-CN estimation was collected from different visits and was derived from buffy coat using the Gentra Puregene Blood Kit (Qiagen). Our analyses were limited to 1,085 individuals with mtDNA-CN data available across all four platforms performed within ARIC: Affymetrix Genome-Wide Human SNP Array 6.0, Illumina HumanExome BeadChip genotyping array, WES and WGS. Eighty-eight percent of our final ARIC participants were African American.

The MESA study recruited 6,814 individuals free of prevalent clinical CVD from 6 US communities across 4 ethnicities. Age range at baseline was 45 to 84 and the baseline exam occurred between 2000 and 2002. DNA for mtDNA-CN analyses was isolated from exam 1 peripheral leukocytes using the Gentra Puregene Blood Kit. Our analyses were restricted to 3,489 white and African American (36%) individuals with mtDNA-CN data available across the three platforms with mtDNA-CN data available at the time of analysis: qPCR, Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina HumanExome BeadChip genotyping array.

All participants provided written informed consent and all centers obtained approval from their respective institutional review boards.

Measurement of mtDNA-CN

qPCR

mtDNA-CN was determined using a multiplexed real time qPCR assay as previously described[30]. Briefly, the cycle threshold (Ct) value of a mitochondrial-specific (*ND1*) and nuclear-specific (*RPPH1*) target were determined in triplicate for each sample. The difference in Ct values (ΔCt) for each replicate represents a raw relative measure of mtDNA-CN. Replicates were removed if they had Ct values for *ND1* > 28, Ct values for *RPPH1* > 5 standard deviations from the mean, or ΔCt values > 3 standard deviations from the mean of the plate. Outlier replicates were identified and excluded for samples with a ΔCt standard deviation > 0.5. The sample was excluded if the ΔCt standard deviation remained > 0.5 after replicate removal. We corrected for an observed linear increase in ΔCt value due to the pipetting order of each replicate via linear regression. The mean ΔCt across all replicates was further adjusted for plate effects as a random effect to represent a raw relative measure of mtDNA-CN.

Microarray

mtDNA-CN was determined using the Genvisis[34] software package for both the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina HumanExome BeadChip genotyping array. A list of high-quality mitochondrial SNPs were hand-curated by employing BLAST to remove SNPs without a perfect match to the annotated mitochondrial location and SNPs with off-target matches longer than 20bp. The probe intensities of the remaining mitochondrial SNPs (25 Affymetrix, 58 Illumina Exome Chip) were determined using quantile sketch normalization (apt-probeset-summarize) as

implemented in the Affymetrix Power Tools software. The median of the normalized intensity, log R ratio (LRR) for all homozygous calls was GC corrected and used as initial estimates of mtDNA-CN for each sample.

Technical covariates such as DNA quality, DNA quantity, and hybridization efficiency were captured via surrogate variable analysis or principal component analysis as previously described[14]. Surrogate variables or principal components were applied to the BLAST filtered, GC corrected LRR of the remaining autosomal SNPs (43,316 Affymetrix, 47,512 Exome Chip).

These autosomal SNPs were selected based on the following quality filters: call rate > 98%, HWE *p* value > 0.00001, PLINK mishap for non-random missingness *p* value > 0.0001, association with sex *p* value > 0.00001, linkage disequilibrium pruning ($r^2 < 0.30$), with maximal spacing between autosomal SNPs of 41.7 kb.

Whole Exome Sequencing

Whole exome capture was performed using Nimblegen's VChrome2.1 (Roche) and sequencing was performed on the Illumina HiSeq 2000. Sequence reads were aligned using Burrows-Wheeler Aligner (BWA)[35] to the hg19 reference genome. Variant calling, and quality control were performed as previously described[36]. mtDNA-CN was calculated using the mitoAnalyzer software package, which determines the observed ratios of sequence coverages between autosomal and mtDNA[37,38].

Due to large batch effects observed in our raw mtDNA-CN calls, alignment summary, insert size, quality score, base distribution, sequencing artifact and quality yield metrics were collected using Picard tools (version 1.87) to take into account differences in

capture efficiency as well as sequencing and alignment quality[39]. Picard sequencing summary metrics to incorporate into our final model were selected through a stepwise backwards elimination model (**Supplementary Table 2.1**).

Whole Genome Sequencing

Whole genome sequencing data was generated at the Baylor College of Medicine Human Genome Sequencing Center using Nano or PCR-free DNA libraries on the Illumina HiSeq 2000. Sequence reads were mapped to the hg19 reference genome using BWA[35]. Variant calling and quality control were performed as previously described[40]. A count for the total number of reads in a sample was scraped from the NCBI sequence read archive using the R package RCurl[41] while reads aligned to the mitochondrial genome were downloaded directly through Samtools (version 1.3.1). A raw measure of mtDNA-CN was calculated as the ratio of mitochondrial reads to the number of total aligned reads. Unlike WES, we did not observe large batch effects in our WGS raw mtDNA-CN calls, obviating the need for adjustment for Picard sequencing summary metrics.

Genotyping and imputation

Genotype calling for the WBC count locus was derived from the Affymetrix Genome-wide Human SNP Array 6.0 in ARIC and MESA. Haplotype phasing for both cohorts was performed using ShapeIt[42] and imputation was performed using IMPUTE2[43]. Genotypes were imputed to the 1000G reference panel (Phase I, version 3). Imputation quality for the Duffy locus lead SNP (rs2814778) was 0.946 and 0.92 in ARIC and MESA, respectively.

DNA extraction method

All DNA used in the DNA extraction comparison were derived from HEK293T cells grown in a single 150T flask to minimize variation due to clonality and cell culture procedures. Extraction were performed with 15 replicates each containing one million cells. mtDNA-CN was determined using qPCR as described previously. To account for the inherent variability in mtDNA-CN estimation, qPCR was run in triplicate.

Silica-base column extraction

We performed a silica-based column extraction using the AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instructions for fewer than 5 x 10^6 cells. Briefly, HEK293T cells were lysed and the subsequent lysate was pipetted directly onto the DNA Allprep spin column for homogenization and DNA binding. The bound DNA was then washed and eluted.

Organic solvent extraction

An aliquot of cells were lysed with 350 µL of RLT Plus Buffer (Qiagen) and one volume of phenol:chloroform:isoamyl alcohol (25:24:1) (PCIAA) was added to the sample and mixed until it turned milky white. The solution was centrifuged and the upper aqueous phase containing DNA was transferred to a separate tube. We proceeded with an ethanol precipitation protocol using 3M sodium acetate to complete the DNA extraction.

Direct cell lysis

Cells were pelleted at 500g for 5 minutes and the supernatant was removed. The cell pellet was agitated in 100 µL of QuickExtract DNA Solution (Lucigen) to disrupt the pellet and placed in a thermocycler for 15 minutes at 68°C followed by 10 minutes at 95°C. The lysate was then centrifuged at 17,000g for 15 minutes to pellet any insoluble inhibitors and the supernatant was transferred to a clean tube. The supernatant containing DNA was finally diluted 1:30 with water to limit the impact of any soluble inhibitors on qPCR.

Statistical analyses

Our final mtDNA-CN phenotype for all measurement techniques is represented as the standardized residuals from a linear model adjusting the raw measure of mtDNA-CN for age, sex, DNA collection center, and technical covariates. Additionally, mtDNA-CN in ARIC was adjusted for WBC count, and the14.9% of individuals with missing WBC data were imputed to the mean. WBC was not available in MESA for the same visit in which the DNA was obtained. As mtDNA-CN was standardized, the effect size estimates are in units of standard deviations, with positive betas corresponding to an increase in mtDNA-CN.

For analyses involving outcomes which also served as covariates in our final phenotype model (age, sex, WBC count), mtDNA-CN was calculated using the full model minus the outcome variable. For example, when exploring the relationship between mtDNA-CN and age, our mtDNA-CN phenotype would represent the standardized residuals from a model controlling for sex, sample collection center, WBC count and any

technical covariates. We would then use this phenotype to explore the association between age and mtDNA-CN such that effect sizes for all comparisons remain in standard deviation units.

Single SNP regression for mtDNA-CN on the WBC count locus was performed in blacks with FAST[44]. In ARIC, mtDNA-CN not adjusted for WBC count was used as the independent variable. Single SNP regression models were additionally adjusted for age, sex, sample collection site, and genotyping PCs.

Cox-proportional hazards regression was used to estimate hazard ratios (HRs) for incident CVD outcomes. Follow-up time was defined from DNA collection through death, loss to follow-up, or study end point (through 2017 in ARIC and 2015 in MESA). Pairwise F-tests were used to test the null hypothesis that the ratio of variances between the DNA extraction methods is equal to one. All analyses were performed using R (version 3.3.3).

RESULTS

The study included 1,085 participants from ARIC with mtDNA-CN data from the Affymetrix 6.0 microarray, the Illumina Exome Chip microarray, WES, and WGS while MESA included 3,489 participants with mtDNA-CN data available from qPCR, the Affymetrix 6.0 microarray, and the Illumina Exome Chip microarray (combined N = 4,574). The mean age of study participants was 61.4 years (ARIC, 57.1 years; MESA 62.7 years), 55.3% of participants were female (n = 2,528), and 46.4% of participants were black (n = 2,124) (**Table 2.1**). While the Affymetrix and Illumina Exome Chip arrays were run in both cohorts, at the time of analysis WES and WGS were unique to ARIC and qPCR was unique to MESA.

mtDNA-CN estimation method comparison

To determine the optimal method for measuring mtDNA-CN, we ranked the performance of each technique based on strength of the association, as measured by *p* values, with the relevant mtDNA-CN correlate (**Supplementary Table 2.2**). Kendall's W tests[45] show significant agreement in rankings across correlates in ARIC ($p = 0.0019$, Kendall's W = 0.79) and MESA ($p = 0.036$, Kendall's W = 0.82) with WGS and the Affymetrix array performing best for each measure in ARIC and MESA, respectively (**Table 2.2**).

To additionally quantify performance, we created a scoring system for each method using negative log transformed *p* values standardized to the least significant method for each correlate. These values were then summed across the correlates for each method to achieve an overall rating of performance (**Supplementary Table 2.3**). These ratings were compared to 1,000 permutations of a random sampling of the standardized and transformed *p* values for each correlate across the different estimation techniques. In ARIC, WGS had a significantly higher performance score compared to all other methods ($p < 0.002$) while the Illumina Exome Chip had a significantly lower score ($p = 0.03$) (**Supplementary Figure 2.1A**). In MESA, Affymetrix had a significantly higher score than qPCR and the Illumina Exome Chip ($p = 0.002$) (**Supplementary Figure 2.1B**). When removing the contribution of WGS in ARIC, the Affymetrix array had a significantly higher score than the Illumina Exome Chip and WES ($p = 0.01$) (**Supplementary Figure 2.1C**).

As WGS and Affymetrix performed similarly, we sought to further parse out their performance by evaluating the 2,746 ARIC samples which contained mtDNA-CN from

both platforms. On average, WGS performed 2.2 orders of magnitude more significantly than the Affymetrix array (**Supplementary Table 2.4**).

DNA extraction comparison

Raw mitochondrial estimates from qPCR were mean-zeroed to the plate average and the mean value across the triplicate plates was used to determine the variance across the 15 replicates for each method (**Fig 2.1**). The variance for our novel Lyse method was significantly lower at 0.02 compared to 0.17 and 0.59 for the PCIAA and Qiagen Kit extractions respectively ($F = 0.13$, $p = 5.44 \times 10^{-4}$; $F = 0.04$, $p = 2.82 \times 10^{-7}$). Additionally, our findings support previous work[33] demonstrating PCIAA had significantly lower variability compared to the Qiagen Kit ($F = 0.29$, $p = 0.03$).

DISCUSSION

We explored several methods for measuring mtDNA-CN in 4,574 self-identified white and black participants from the ARIC and MESA studies. We found mtDNA-CN estimated from WGS read counts and Affymetrix Genome-Wide Human SNP Array 6.0 probe intensities was more significantly associated with known mtDNA-CN correlates compared to mtDNA-CN estimated from WES, qPCR and the Illumina HumanExome BeadChip. When observing the relative performance of these methods, mtDNA-CN calculated from either WGS or Affymetrix array are, respectively, 5.6 and 5.4 orders of magnitude more significant than the current gold standard of qPCR (**Figure 2.2**). These results are not limited to significance as we see similar trends when exploring effect size estimates (**Figure 2.3**). For example, when looking at incident CVD, mtDNA-CN

measured from WGS observes a substantial HR of 0.63 (0.54 – 0.74) where as mtDNA-CN measured from qPCR only has a HR of 0.93 (0.82 – 1.05), a marked difference. As a result, when exploring the relationship between mtDNA-CN and a trait of interest, on average one could expect a result 5.6 orders of magnitude less significant and 6 times less extreme when using mtDNA-CN estimated from qPCR data as opposed to WGS.

Interestingly, mtDNA-CN measured from two seemingly similar microarray platforms differed drastically (**Supplementary Figure 2.2**). However, this finding is unsurprising when exploring the underlying biochemistry of sample preparation for each microarray platform. While the Affymetrix protocol starts with two restriction enzyme digests prior to whole genome amplification (WGA), the Illumina Exome Chip requires WGA with a processive polymerase prior to sonication. As a result, the mitochondrial genome undergoes rolling circle amplification which occurs at a significantly faster rate than linear WGA[46].

Lower mtDNA-CN has been found to be associated with an increased incidence for several diseases, including end stage renal disease, type 2 diabetes, and non-alcoholic fatty liver disease[47–49]. However, such studies have relied on mtDNA-CN estimated from qPCR data. Our findings suggest much of the current literature may be severely underestimating disease associations with mtDNA-CN as well as its potential as a predictor of disease outcomes. Despite this, at <$2 per sample qPCR may remain the principal method for measuring mtDNA-CN due to the prohibitive costs of WGS. As a result, it may be time for the field to start exploring other low cost methods, such as digital droplet PCR, which may improve upon the accuracy of qPCR[50,51].

We additionally showed DNA extraction method affects mtDNA-CN estimate reproducibility with copy number measured directly from cell lysate significantly outperforming silica-based column extraction and organic solvent extraction. Although several other studies have explored the impact of DNA isolation protocol on mtDNA-CN estimation[33,52,53], to our knowledge, this is the first study to interrogate the possibility of measuring mtDNA-CN directly from cell lysate. In addition to the superior performance of direct cell lysis, this method is cheaper and has less hands-on time than PCIAA or Qiagen Kit extractions. However, the authors recognize DNA from cell lysate has less downstream utility than traditional DNA extraction procedures potentially limiting its adoption within the mtDNA-CN field when sample availability is limited. Furthermore, it is important to note the various DNA extraction methods resulted in significantly different mtDNA-CN estimates ($p = 3.56\text{x}10^{-11}$, 0.02, $2.85\text{x}10^{-7}$ for Lyse:PCIAA, Lyse:Qiagen Kit, and PCIAA:Qiagen Kit respectively). As such, when choosing an extraction method, it is important to remain consistent across the study.

In conclusion, our study demonstrates mtDNA-CN calculated from WGS reads or Affymetrix microarray probe intensities significantly improves upon the current gold standard method of qPCR. Furthermore, we show direct cell lysis introduces less variability to mtDNA-CN estimates than popular DNA extraction methods. Despite the relative infancy of using mtDNA-CN as a novel risk marker, these findings highlight the need for the field to adapt to current technologies to ensure disease and trait associations are fully realized with a move toward more accurate microarray and WGS methods. Furthermore, due to the prevalence of qPCR in the literature, the authors recommend re-analyzing trait associations as more WGS data becomes available from large initiatives such as TOPMed.

FIGURES AND TABLES

Table 2.1. Participant characteristics

Participant Characteristics	ARIC	MESA
n	1,085	3,489
Sex (female)	672 (61.9)	1,856 (53.2)
Race (black)	958 (88.3)	1,226 (35.1)
Age	57.1 ± 5.9	62.7 ± 10.2
WBC count (10^3/µl)	5.8 ± 1.7	NA
Incident CVD	174 (16.0)	270 (7.7)

Values are number (%) or mean ± SD
Abbreviations: SD, standard deviation; WBC, white blood cell; CVD, cardiovascular disease

Table 2.2. Performance rankings for mtDNA-CN estimation methods

Cohort	Assay	Age	Sex	WBC	Duffy locus*	Incident CVD	Mean Rank	Kendall's W *p* value
ARIC	Exome	2	4	3	4	4	3.4	0.001
	Affy	3	2	2	2	2	2.2	
	WES	4	3	4	3	3	3.4	
	WGS	1	1	1	1	1	1	
MESA	Exome	2	3	NA	2.5	3	2.625	0.03
	Affy	1	1	NA	1	1	1	
	qPCR	3	2	NA	2.5	2	2.375	

*Duffy locus associations were performed in blacks only

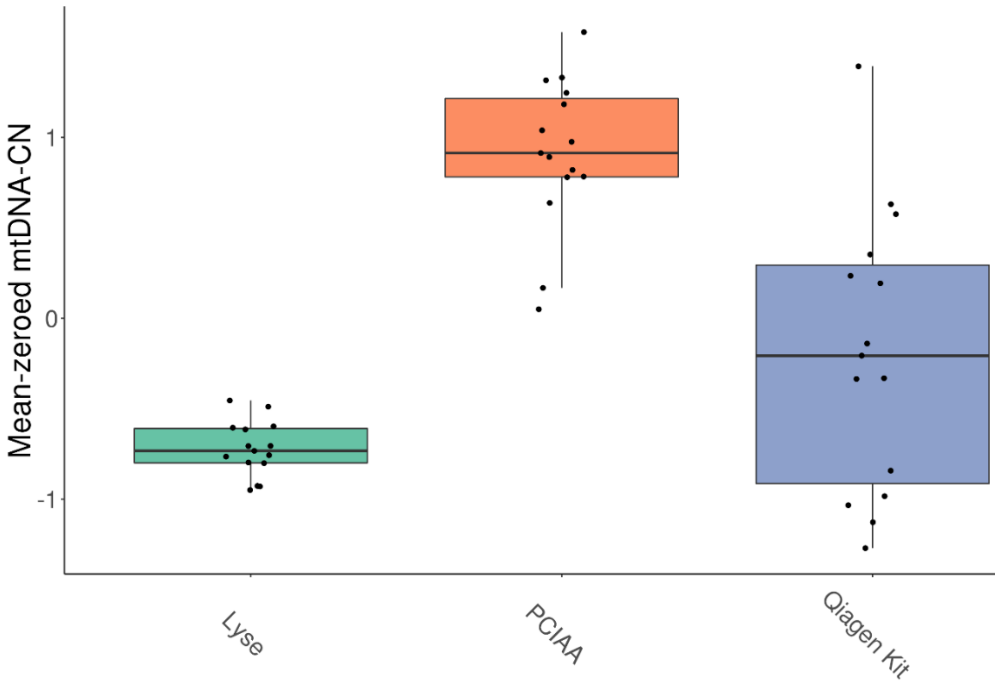


Figure 2.1. mtDNA-CN measured across DNA extraction methods. mtDNA-CN measured by qPCR was mean-zeroed and averaged across three runs for Lyse, PCIAA and Qiagen Kit DNA extractions. Variance for Lyse, PCIAA and Qiagen Kit are 0.02, 0.17 and 0.59 respectively. PCIAA, phenol:chloroform:isoamyl alcohol.

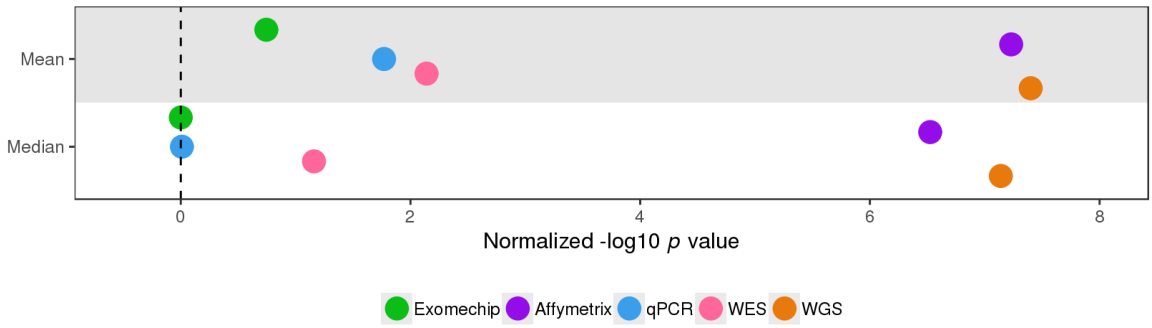
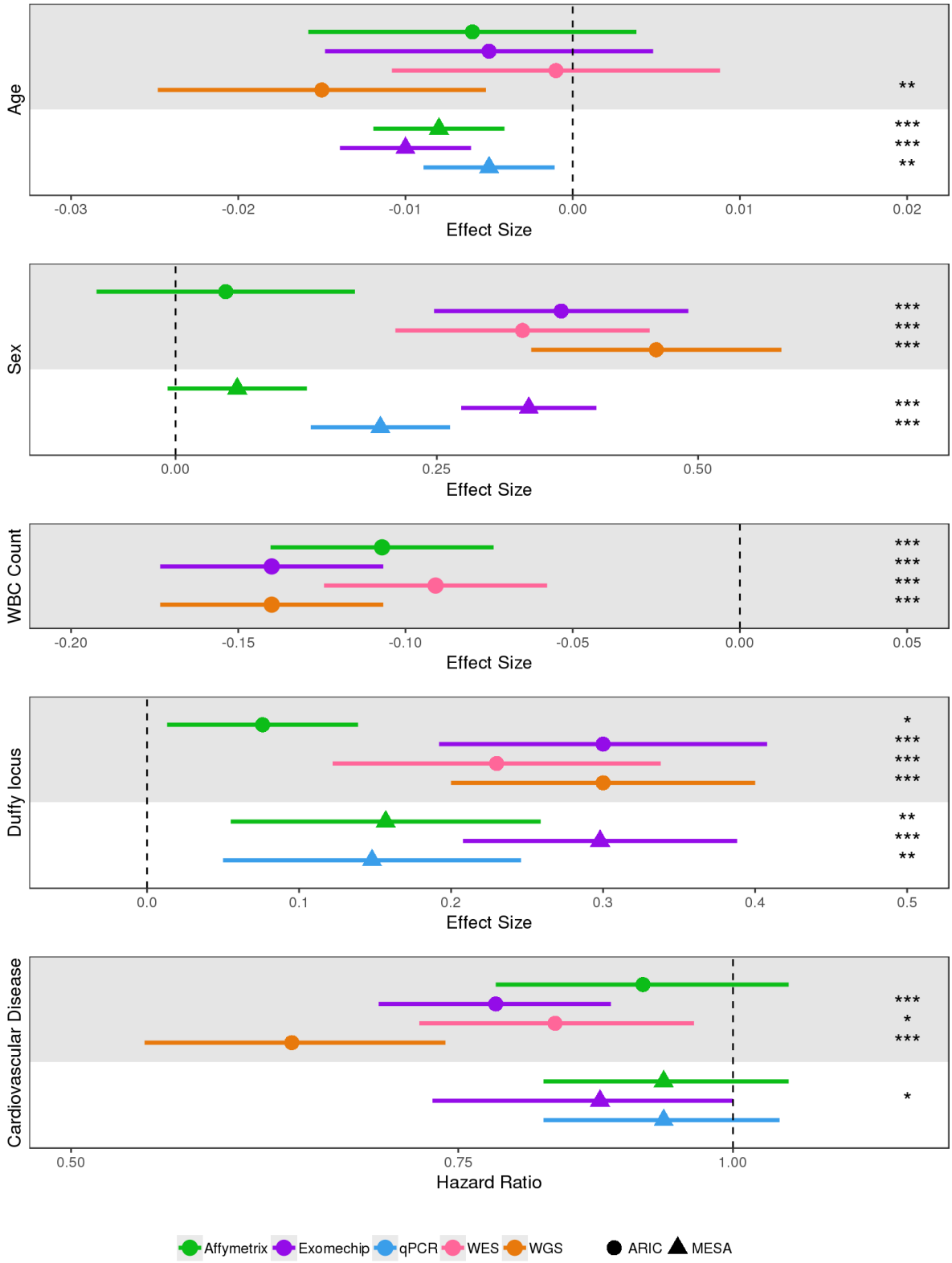


Figure 2.2. Relative overall performance of mtDNA-CN estimation methods. Overall performance for each method scored as mean or median of the negative log-transformed *p* value across all correlates normalized to the least significant method of each correlate. For ExomeChip and Affymetrix, the mean value across both cohorts was used as the final measure of performance.

Figure 2.3. Effect size and Hazard Ratio estimates for mtDNA-CN with known correlates. Data points and their corresponding 95% confidence intervals represent the effect size or Hazard Ratio estimates for mtDNA-CN with Age, Sex, WBC count, Duffy locus, and incident Cardiovascular Disease. Effect size estimates are in standard deviation units. The significance of each estimate is represented as '*' for $P < 0.05$, '**' for $P < 0.01$, and '***' for $P < 0.001$. WBC, white blood cell.


Age
**
***
***
**
-0.03
-0.02
-0.01
0.00
0.01
0.02
Effect Size
Sex
***
***
***
***
***
0.00
0.25
0.50
Effect Size
WBC Count
***
***
***
***
-0.20
-0.15
-0.10
-0.05
0.00
0.05
Effect Size
Duffy locus
*
***
***
***
**
***
**
0.0
0.1
0.2
0.3
0.4
0.5
Effect Size
Cardiovascular Disease
***
*
***
*
0.50
0.75
1.00
Hazard Ratio
Affymetrix
Exomechip
qPCR
WES
WGS
ARIC
MESA

SUPPLEMENTARY MATERIALS

Supplementary Table 2.1. Picard sequencing summary metrics definitions

Category	Picard Metric	Measurement
First of pair	PCT_PF_READS	Fraction of reads which pass Illumina's filter
	PF_MISMATCH_RATE	Rate of bases mismatching the reference for all bases aligned to the reference sequence
	PF_HQ_ERROR_RATE	Fraction of bases from reads with mapping quality ≥ Q20 which mismatch the reference
	PF_INDEL_RATE	Number of indel events per 100 aligned bases
	MEAN_READ_LENGTH	Mean read length
	PCT_CHIMERAS	Fraction of reads that map outside of a maximum insert size or have two ends mapping to different chromosomes
Second of pair	PF_MISMATCH_RATE	Rate of bases mismatching the reference for all bases aligned to the reference sequence
	PF_HQ_ERROR_RATE	Fraction of bases from reads with mapping quality ≥ Q20 which mismatch the reference
	PF_INDEL_RATE	Number of indel events per 100 aligned bases
	PCT_CHIMERAS	Fraction of reads that map outside of a maximum insert size or have two ends mapping to different chromosomes
Pair	PF_MISMATCH_RATE	Rate of bases mismatching the reference for all bases aligned to the reference sequence
	PF_HQ_ERROR_RATE	Fraction of bases from reads with mapping quality ≥ Q20 which mismatch the reference
	PF_INDEL_RATE	Number of indel events per 100 aligned bases
	STRAND_BALANCE	Ratio of reads which pass Illumina's filter aligned to the positive strand versus those aligned to the negative strand

Supplementary Table 2.2. Associations of known correlates with mtDNA-CN estimation Platforms

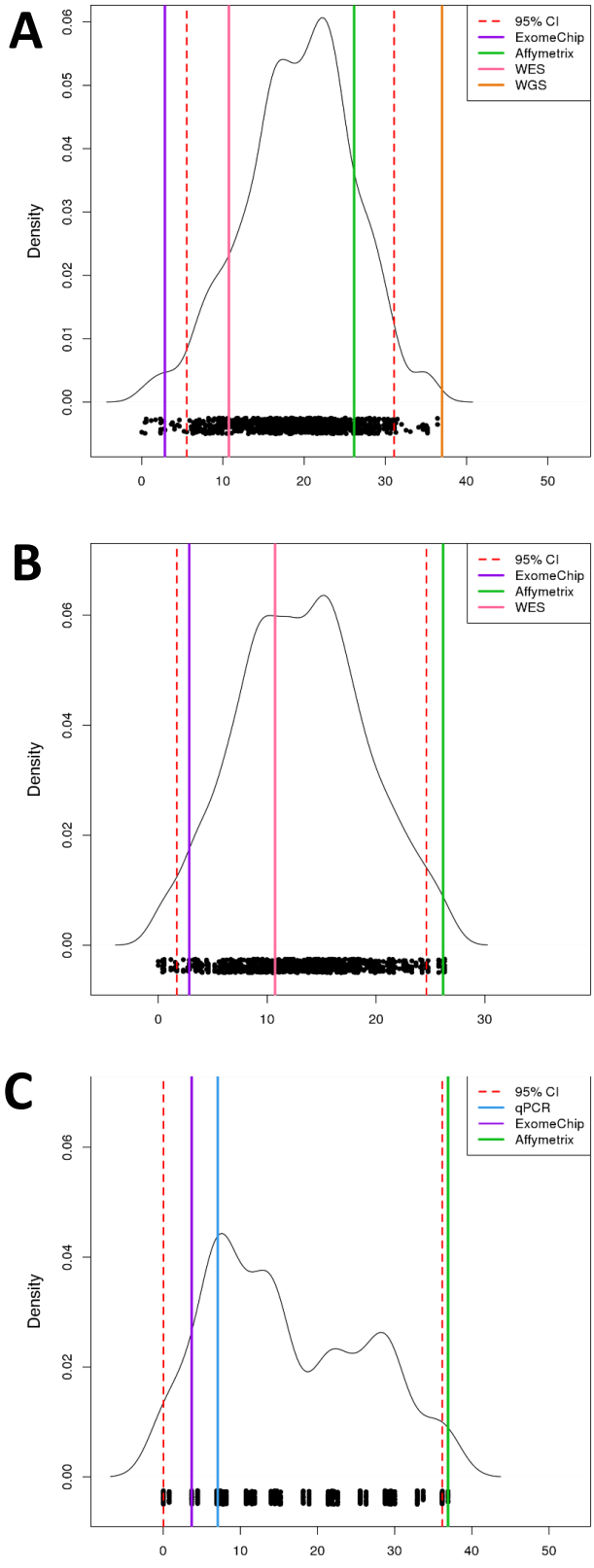
Cohort	Assay	Age			Sex			WBC			Duffy locus*			CVD		
		p value	Beta	SE	*p* value	Beta	SE	*p* value	Beta	SE	*p* value	Beta	SE	*p* value	HR	95% CI
ARIC	Exome	0.23	-0.006	0.005	0.44	0.05	0.06	9.04E-10	-0.107	0.02	0.02	0.08	0.03	0.21	0.91	0.78 - 1.06
	Affy	0.33	-0.005	0.005	2.69E-09	0.37	0.06	7.09E-16	-0.14	0.02	2.57E-08	0.30	0.06	1.18E-04	0.78	0.69 - 0.88
	WES	0.82	-0.001	0.005	9.02E-08	0.33	0.06	1.81E-07	-0.091	0.02	2.65E-05	0.23	0.06	0.01	0.83	0.72 - 0.96
	WGS	0.004	-0.02	0.005	9.95E-14	0.46	0.06	6.97E-16	-0.14	0.02	7.14E-09	0.30	0.05	1.47E-08	0.63	0.54 - 0.74
MESA	Exome	3.77E-07	-0.008	0.002	0.08	0.06	0.03	NA	NA	NA	0.003	0.16	0.05	0.26	0.93	0.82 - 1.06
	Affy	2.21E-10	-0.01	0.002	1.16E-23	0.34	0.03	NA	NA	NA	1.19E-10	0.30	0.05	0.04	0.87	0.73 - 1.00
	qPCR	0.002	-0.005	0.002	7.09E-09	0.20	0.03	NA	NA	NA	0.003	0.15	0.05	0.25	0.93	0.82 - 1.05

*Duffy locus associations were performed in blacks only

Supplementary Table 2.3. Relative performance of methods as rated by standardized –log *p* values

Cohort	Assay	Age	Sex	WBC	Duffy locus*	Incident CVD	mean	median
ARIC	Exome	0.55	0	2.30	0	0	0.57	0
	Affy	0.40	8.21	8.41	5.89	3.24	5.23	5.89
	WES	0	6.69	0	2.88	1.16	2.14	1.16
	WGS	2.34	12.65	8.41	6.45	7.14	7.40	7.14
MESA	Exome	3.69	0	NA	0	0	0.92	0
	Affy	6.92	21.84	NA	7.40	0.77	9.23	7.16
	qPCR	0	7.05	NA	0	0.02	1.77	0.01

*Duffy locus associations were performed in blacks only

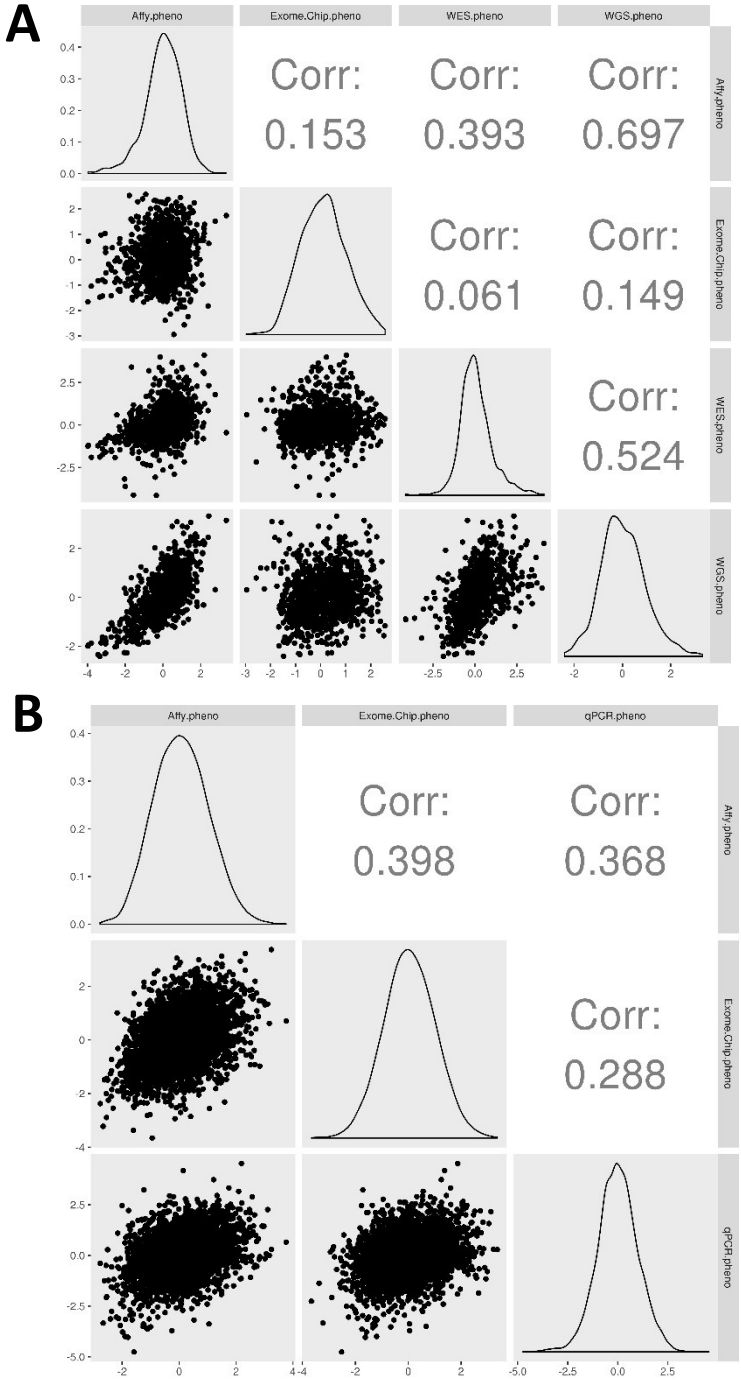


Supplementary Figure 2.1. Permutation test for mtDNA-CN estimation method performance Performance scored as sum of negative log-transformed *p* value for each method across all correlates normalized to the least significant method of each correlate. Compared to 1,000 permutations of the sums of randomly selected normalized and transformed *p* values. ARIC (A), MESA (B), ARIC without WGS (C).

Supplementary Table 2.4. Relative performance of WGS and Affymetrix as rated by standardized –log *p* values

Cohort	Assay	Age	Sex	WBC	Duffy locus*	Incident CVD	mean	median
ARIC	Affy	0	0	0	0	0	0	0
	WGS	1.04	0.63	3.82	2.26	3.29	2.21	2.26

*Duffy locus associations were performed in blacks only



Supplementary Figure 2.2. Phenotype correlation plots. ARIC (A) and MESA (B)

Chapter 3 :
A genome-wide association study of mitochondrial DNA copy number in 79,444 individuals from the CHARGE consortium and UK Biobank

INTRODUCTION

Unlike the nuclear genome, a large amount of variation exists in the number of copies of mtDNA present within cells, tissues, and individuals. The relative copy number of mtDNA (mtDNA-CN) has been shown to directly correlate with oxidative stress, energy reserves, and mitochondrial membrane potential[13]. As a minimally invasive proxy measure of mitochondrial dysfunction[11], mtDNA-CN has been previously associated with several disease states including frailty[30], cardiovascular disease (CVD)[14], chronic kidney disease[54], neurodegeneration[23,24], and cancer[25].

Although the comprehensive mechanism through which mtDNA-CN is modulated is largely unknown[15,55], several genes have been shown to influence mtDNA-CN. Proteins within the mtDNA replication machinery directly modulate mtDNA-CN, including those of the mitochondrial polymerase, *POLG* and *POLG2*,[56,57] as well as the mitochondrial DNA helicase, *C10orf2*, and the mitochondrial single-stranded binding protein, *mtSSB*[58]. *TFAM* initiates mtDNA replication and is a common target for *in vivo* studies looking to modulate mtDNA-CN[32,59]. Furthermore, genes which maintain proper mitochondrial nucleotide supply including *DGUOK* and *TK2* have also been shown to regulate mtDNA-CN[60,61]. However, the impact of these genes on mtDNA-CN were discovered through Mendelian mitochondrial diseases and do not sufficiently represent the large number of loci which likely control a complex phenotype such as mtDNA-CN.

To date, several genome-wide association studies (GWAS) of mtDNA-CN have been published, unfortunately due to their limited sample sizes few genome-wide significant loci have been identified[32,62,63]. In a GWAS of 10,442 Han Chinese women

recruited through the China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) consortium, Cai et al. found two putative loci which control mtDNA-CN[32]. Their first hit occurred within the 3' region of *TFAM*, a gene already known to regulate mtDNA replication, while their second hit occurred within the first intron of *CDK6*, a gene which has not been previously implicated in mtDNA-CN. These findings were further replicated by 1,753 individuals of the Avon Longitudinal Study of Parent and Children (ALSPAC) study. In spite of these findings, given the polygenic nature of many complex phenotypes, the field is most likely only scratching the surface in terms of genes which contribute to mtDNA-CN variation.

In this study we leverage mtDNA-CN measured in 79,444 individuals across six prospective cohorts and four ethnicities to perform the largest mtDNA-CN GWAS to date. We additionally validate loci from previous studies and perform gene ontology and pathway analysis to identify novel networks which may regulate mtDNA-CN.

METHODS

All participants provided written informed consent and all centers obtained approval from their respective institutional review boards. Detailed cohort descriptions and ethics statements available within the Supplementary Materials.

Estimation of Mitochondrial DNA Copy Number

qPCR

mtDNA-CN was determined using a multiplexed real time qPCR assay as previously described[64]. Briefly, the cycle threshold (Ct) value of a nuclear-specific (*RPPH1*) and mitochondrial-specific (*ND1*) probe were measured in triplicate for each sample. In addition to plate effects, we observed a linear increase in ΔCt due to the pipetting order of each replicate. We corrected for these effects by mixed linear regression whereby pipetting order was a fixed effect and plate was a random effect.

Microarray

Microarray probe intensities were used to estimate mtDNA-CN using the Genvisis software package as previously described[14,34,64]. Briefly, the mitochondrial probe intensities were determined using quantile sketch normalization (apt-probeset-summarize) as implemented in the Affymetrix Power Tools software[65]. The median of the log R ratio normalized probe intensity for all homozygous calls was GC corrected and used as initial estimates of mtDNA-CN. Technical artifacts such as DNA quality and DNA quantity were captured via surrogate variable analysis or principal component analysis.

Whole Exome Sequencing

Detailed methods for estimating mtDNA-CN from UK Biobank whole exome sequencing read counts can be found in the Supplementary Material. Briefly, a linear regression model was used to adjust mtDNA read count for total, unknown, decoy1 and decoy2 read counts. Residuals from this model were then further adjusted for age, sex,

platelet count and neutrophil count to represent a final measure of mtDNA-CN (**Supplementary Table 3.1**).

Whole Genome Sequencing

Whole genome sequencing read counts were used to estimate mtDNA-CN as previously described[64]. Briefly, the total amount of reads in a sample was scraped from the NCBI sequence read archive. Mitochondrial reads were downloaded directly through Samtools (1.3.1). A ratio of mitochondrial reads to total aligned reads was used as a raw measure of mtDNA-CN.

Adjusting for covariates

The final mtDNA-CN phenotype is represented as the standardized residuals (mean = 0, standard deviation = 1) from a linear model adjusting for known influences on mtDNA-CN including age, sex, and DNA collection site. Estimates from ARIC, SHIP, and CHS were additionally adjusted for white blood cell count. Details on covariate adjustment for UKB is available within the Supplementary Material.

Genome-wide association study

Each study performed regression analysis with mtDNA-CN as the dependent variable adjusting for age, sex, white blood cell count (if applicable), and cohort-specific covariates (principal components, DNA collection site, family structure). Race stratified meta-analyses were performed with Metasoft[66] with random effects models to control for unobserved heterogeneity due to differences in mtDNA-CN estimation method between

studies. Similarly, trans-ethnic meta-analyses were conducted with a random effects model via Metasoft. Effect size estimates for SNPs were calculated via standard inverse variance-weighted meta-analysis using cohort summary statistics.

GO/KEGG

Gene-based testing with 50 kb windows around each gene was performed with VEGAS2[67]. Gene Ontology (GO)[68,69] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[70,71] analyses were performed using gene-based testing results. We performed 9 stepwise cutoffs of input genes ranging from 100 to 500 as well as 1,000 permutation tests to ensure robustness of results.

RESULTS

Sample characteristics

The study included 79,444 individuals (54.9% female) across four ethnicities (6.0% Black, 1.0% Chinese, 1.5% Hispanic) (**Table 3.1**). A majority of the data originated from the UK Biobank (56.4%) and consisted of mtDNA-CN derived from WES. mtDNA-CN estimated from the Affymetrix array consisted of 39.0% of the data while qPCR-derived mtDNA-CN consisted of only 4.6% of the data.

Genome-wide association study

Trans-ethnic meta-analysis reveals four novel genome wide significant loci (**Figure 3.1**, **Table 3.2**, **Supplementary Figure 3.1**). The lead SNP of the *NDUFV3* locus (rs4148974) occurs within exon 3 of *NDUFV3* and the lead SNP of the *REEP3* locus

(rs7895549) occurs within the 3rd intron of *REEP3* while the other two loci are intergenic. Manhattan/QQ plots for ethnicity specific GWAS available in **Supplementary Figures 3.2A-D**, **3.3A-D**.

In an attempt to validate previous findings, we validate the *TFAM* (rs11006126) and *CDK6* (rs445) loci from the Cai et al. GWAS (**Table 3.3**). The SNP rs445 was significantly associated with mtDNA-CN in our GWAS (p value = 1.33 x 10^{-4}) while rs11006126 was only nominally associated with mtDNA-CN (p value = 0.059). Our effect size estimates are considerably smaller than those found in the CONVERGE and ALSPAC studies, however this is most likely due to different standardizations of the mtDNA-CN phenotype. Whereas the Cai et al. paper transformed mtDNA-CN to normality by quantile-normalization after adjusting for covariates, we standardized our phenotype such that it was represented in standard deviation units (mean = 0, sd = 1). While both methods are sound, our approach allows our results to be interpreted in terms of units in reference to the overall population making it easier to interpret.

GO/KEGG

Gene-based testing and subsequent pathway and biological processes analyses through GO and KEGG reveal a potential link to *alkaline phosphatase activity* (p value = 1.38 x 10^{-5}), the only pathway to remain significant after permutation testing (permutation p value cutoff = 2.11 x 10^{-4}) (**Table 3.4**). Although not significant, GO identified multiple pathways involved with guanosine including *dGTP metabolic process* (p = 1.36 x 10^{-3}), *GMP metabolic process* (p = 1.39 x 10^{-3}), and *guanosine-containing compound metabolic process* (p = 2.85 x 10^{-3}). KEGG pathway analysis revealed important pathways for proper

mitochondrial metabolism including *thiamine metabolism* (p = 4.88 x 10^{-3}) and *folate biosynthesis* (p = 8.73 x 10^{-3}). Both GO and KEGG additionally identified processes involved directly in mitochondrial energy production in *proton-transporting two-sector ATPase complex, proton-transporting domain* (p = 2.83 x 10^{-3}) and *oxidative phosphorylation* (p = 0.033).

DISCUSSION

We performed a trans-ethnic GWAS on mtDNA-CN in 79,444 individuals from the CHARGE consortium and the UK Biobank. We identified four novel loci as well as validated previous loci implicated in mtDNA-CN modulation from a previous GWAS of 10,000 Chinese women. Furthermore, genes near GWAS loci were implicated in alkaline phosphatase activity, a biological process not previously implicated in mitochondrial function or mtDNA-CN maintenance.

Of our genome-wide significant hits, the exonic signal found within *NDUFV3* is of particular interest as this gene produces one of the three subunits of the flavoprotein fraction of NADH dehydrogenase (complex I). Although the function of NDUFV3 is unknown, the flavoprotein fraction plays a catalytic role in the oxidation of NADH, directly implicating it in mitochondrial function[72]. Additionally, *CCDC71L* has been previously linked to carotid media thickness, highlighting a potential link to CVD[73]. Unlike *NDUFV3* and *CCDC71L*, genes located within the *ZNRF4* and *REEP3* loci have not been previously implicated in mitochondrial function or CVD.

While alkaline phosphatase has been found to be active within the mitochondria[74], little is known about its overall role in mitochondrial function. Of interest, elevated levels

of alkaline phosphatase in the blood are often the result of liver disease, which has been previously associated with mtDNA-CN[49,75,76]. Although not significant after permutation testing, Gene ontology additionally identified several processes involved with the metabolism of guanosine. Previous research in rats has identified dGTP represents about 90% of the total dNTP pool from certain tissues[77,78]. This dNTP imbalance could lead to insertion or proofreading errors during mtDNA replication[79]. Additionally, the oxidizing environment of the mitochondrion could lead to oxidative mutagenesis due to the incorporation of 8-oxo-dTP[78]. Regulation of mtDNA-CN through guanosine metabolism could further link copy number to mitochondrial function as mtDNA mutations have been previously linked to oxidative stress and mitochondrial dysfunction[80,81].

Although they did not reach the permutation *p* value threshold, thiamine metabolism and folate biosynthesis were identified as two of the top hits in KEGG pathway analysis. Thiamine pyrophosphate, a thiamine derivative, is an essential cofactor for pyruvate, α-ketoglutarate, and branched-chain ketoacid dehydrogenases[82], and thus critical for mitochondrial energy production. Folate deficiency mediated through altered folate metabolism has been previously linked to improper mitochondrial translation[83] , mtDNA instability[84,85], and thus mitochondrial dysfunction. Furthermore, as the one carbon donor, folate metabolism is critical to both DNA[86] and histone methylation[87], which was also identified in our GO analyses.

In conclusion, we identify and validate several putative modulators of mtDNA-CN which warrant further functional follow up. Additionally, our GO and KEGG analyses highlight that despite our large sample size, many regulators of mtDNA-CN remain below

the genome-wide significance threshold. As such, we propose further investigation into the genetic regulators of mtDNA-CN as more data comes available.

FIGURES AND TABLES

Table 3.1. Sample characteristics

Participant Characteristics	ARIC	ARIC	CHS	FHS	MESA	ROSMAP	SHIP	UKBiobank
n	2,921	8,685	3,670	7,797	5,916	1,708	3,942	44,805
mtDNA-CN platform	WGS	Affymetrix	qPCR	Affymetrix	Affymetrix	Affymetrix	Affymetrix	WES
Sex (female)	1,697 (58.1)	4,721 (54.4)	2,230 (60.8)	4,267 (54.7)	3,096 (52.3)	1,182 (69.2)	2,010 (51.0)	24,315 (54.3)
Ethnicity								
White	1,453 (49.7)	7,552 (87.0)	2,938 (80.1)	7,797 (100)	2,523 (42.6)	1,708 (100)	3,942 (100)	44,805 (100)
Black	1468 (50.3)	1,133 (13.0)	732 (19.9)		1,437 (24.3)			
Chinese					775 (13.1)			
Hispanic					1,181 (20.0)			
Age	57.6 ± 6.0	58.0 ± 5.9	72.2 ± 5.3	52.4 ± 16.2	62.4 ± 10.3	78.5 ± 7.5	49.6 ± 16.2	56.8 ± 7.9
WBC count (10^3/μl)	6.0 ± 1.7	6.2 ± 1.7	6.2 ± 2.1				6.7 ± 2.0	7.0 ± 1.7

Values are number (%) or mean ± SD. Abbreviations: WGS, whole genome sequencing; WES, whole exome sequencing; WBC, white blood cell count

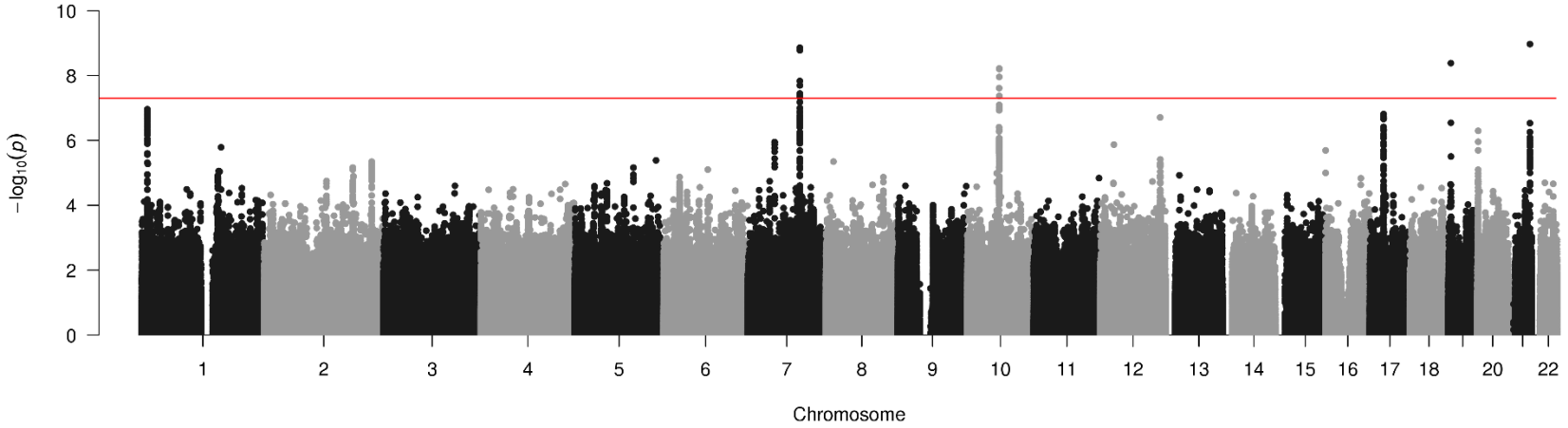


Figure 3.1. Trans-ethnic meta-analysis of 79,444 individuals reveals four novel genome-wide significant loci. Random effects meta-analysis to account for differences due to ethnicity and mtDNA-CN estimation platform. Filters: Imputation quality > 0.6, MAF > 0.005, SNP had to be present in at least half of the cohorts.

Table 3.2. Summary statistics for genome-wide significant hits

SNP	Chr	Pos	Coded Allele	Nearest Gene	Beta	SE	P
rs4148974	21	44323720	T	*NDUFV3* (exon 3)	-0.067	0.011	1.07×10^{-9}
rs342259	7	106351452	T	*CCDC71L*	0.034	0.006	1.39×10^{-9}
rs407949	19	5498934	G	*ZNRF4*	-0.055	0.009	4.15×10^{-9}
rs7895549	10	65357438	A	*REEP3* (intron 3)	-0.030	0.005	6.04×10^{-9}

Abbreviations: SNP, single nucleotide polymorphism; Chr, chromosome; Pos, genomic position; Beta, effect size estimate; SE, standard error; P, *p* value

Table 3.3. Replication of Cai et al. mtDNA-CN GWAS

Gene (SNP)	Cohort	Beta	SE	P
CDK6 (rs445)	Converge	0.119	0.015	4.57×10^{-16}
	ALSPAC	0.110	0.057	0.021
	CHARGE + UKB	0.035	0.009	1.33×10^{-4}
TFAM (rs11006126)	Converge	-0.195	0.018	1.17×10^{-27}
	ALSPAC	-0.179	0.047	1.53×10^{-4}
	CHARGE + UKB	-0.020	0.010	0.059

Abbreviations: SNP, single nucleotide polymorphism; Beta, effect size estimate; SE, standard error; P, *p* value

Table 3.4. GO and KEGG pathway and biological process analysis

	Pathway ID	Name of Pathway or Biological Process	Number of Genes in Pathway ID	Number of submitted Genes in Pathway ID	P
GO	0004035	alkaline phosphatase activity	4	3	1.38×10^{-5}
	0080182	histone H3-K4 trimethylation	13	3	8.94×10^{-4}
	0046070	dGTP metabolic process	4	2	1.36×10^{-3}
	0046037	GMP metabolic process	15	3	1.39×10^{-3}
	0097381	photoreceptor disc membrane	17	3	2.03×10^{-3}
	0033177	proton-transporting two-sector ATPase complex, proton-transporting domain	19	3	2.83×10^{-3}
	1901068	guanosine-containing compound metabolic process	39	4	2.85×10^{-3}
KEGG	hsa04810	regulation of actin cytoskeleton	175	6	1.08×10^{-3}
	hsa00730	thiamine metabolism	15	2	4.48×10^{-3}
	hsa00790	folate biosynthesis	21	2	8.73×10^{-3}
	hsa03410	base excision repair	27	2	0.014
	hsa04670	leukocyte transendothelial migration	98	3	0.028
	hsa00190	oxidative phosphorylation	104	3	0.033
	hsa05030	cocaine addiction	44	2	0.036

Top 7 hits for GO and KEGG analyses. Cutoff for GO analysis: 300 genes. Cutoff for KEGG analysis: 150 genes. *p* value for over-representation of submitted genes within the pathway. Abbreviations: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; P, *p* value

SUPPLEMENTARY MATERIAL

UK Biobank whole exome sequencing mtDNA-CN estimation

We started with 49,997 Exome SPB CRAM files (version Jul 2018) downloaded from the UKB data repository, and used Samtools (ver1.9) to extract read summary statistics ('idxstats' command). Linear regression models were used to adjust for total DNA and potential technical artifacts. Specifically, we used 10-fold cross validation for variable selection, using the 'leaps' R package (version 3.0), with an initial model with chrMT read count as the dependent variable, and 'Total', 'Mapped', 'unknown', 'random', 'decoy1' and 'decoy2' read counts as the independent variables. For each of the independent variables, we included a natural spline with df=4 to allow for non-linear effects. The independent variables 'Total', 'unknown', 'decoy1' and 'decoy2' read counts were selected. We then increased the natural spline df to 15, and then used backward selection to reduction model complexity, requiring P<0.005 to keep a term in the model. The final regression model residuals were generated with the following R (version 3.6.0) code:

resid.mtDNA= residuals(lm(chrMT ~ ns(Total,df=3) + ns(unknown,df=4) + ns(decoy1,df=7) + decoy2))

Residuals from this model were then adjusted for age, with a natural spline df=2, and sex. In addition, we performed 10-fold cross validation as described above to identify cell type counts to include as covariates in the regression model. The initial model included Platelets, Nucleated RBC (yes/no), and log+1 transformations of WBC, RBC, Lymphocytes,

Monocytes, Neutrophils, Eosinophils, Basophils, and Nucleated RBC (continuous) counts. With the exception of both Nucleated RBC measures, natural splines were included with df=3. The final regression model residuals were generated with the following R (version 3.6.0) code:

final.mtDNA = residuals(lm(resid.mtDNA ~ sex + ns(age,df=2) + ns(Platelet,df=3) + ns(Neutrophil,df=2)))

For these analyses, cell type count outliers were removed based on the following criteria:

Log(WBC) ≤ 1.25 or ≥ 3
Log(RBC) ≤ 1.4 or ≥ 2
Platelet ≤10 or ≥ 500
Log(Lymphocytes) ≤ 0.10 or ≥ 2
Log(Monocytes) ≥ 0.9
Log(Neutrophils) ≤ 0.75 or ≥ 2.75
Log(Eosinophils) ≥ 0.75
Log(Basophils) ≥ 0.45

Cohort descriptions

ARIC

The Atherosclerosis Risk in Communities study (ARIC) recruited 15,792 individuals between 1987 and 1989 aged 45 to 65 years from 4 US communities[88]. DNA for mtDNA-CN estimation was collected from different visits and was derived from buffy coat using the Gentra Puregene Blood Kit (Qiagen). Whole genome sequencing (WGS) data was generated at the Baylor College of Medicine Human Genome Sequencing Center using Nano or PCR-free DNA libraries on the Illumina HiSeq 2000. Genotyping for the Affymetrix Genome-Wide 6.0 Human SNP Array 6.0 was performed in accordance to the manufacturers protocol and genotypes were called using Birdseed (version 2).

With the availability of two different mtDNA-CN estimation platforms, and previous researching highlighting WGS outperforms the Affymetrix array in mtDNA-CN estimation[64], mtDNA-CN was called first from the 2,923 individuals with WGS data. Copy number was additionally estimated in 8,687 from the Affymetrix genotyping array for individuals without WGS data. WGS and Affymetrix mtDNA-CN batches were treated as separate cohorts due to differences in population distribution. Missing white blood cell count data (14.7%) was imputed to the mean.

CHS

The Cardiovascular Health Study (CHS) is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers[89]. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons were enrolled for a total sample of 5,888.

Blood samples were drawn from all participants at their baseline examination and DNA was subsequently extracted from available samples. Genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai among CHS participants who consented to genetic testing and had DNA available using the Illumina 370CNV BeadChip system (for European ancestry participants, in 2007) or the Illumina HumanOmni1-Quad_v1 BeadChip system (for African-American participants, in 2010).

FHS

In 2007, the Framingham Heart Study (FHS) entered a new phase with the conduct of genotyping for the FHS SHARe project, for which genotyping was conducted using approximately 550,000 SNPs (Affymetrix 500K mapping array plus Affymetrix 50K supplemental array) in over 9,300 subjects from the three generations of subjects (including over 1500 families).

MESA

The MESA study recruited 6,814 individuals from 6 US communities free of prevalent clinical CVD across 4 ethnicities. Age range at baseline was 45 to 84 and the baseline exam occurred between 2000 and 2002. Affymetrix Genome-Wide Human SNP Array 6.0 genotype data was available for 8,227 unique individuals within the MESA cohort. DNA derived from MESA Family, a subset of MESA, originated from cell lines and was excluded resulting in a final sample size of 5,916. DNA for mtDNA-CN analyses was isolated from exam 1 peripheral leukocytes using the Gentra Puregene Blood Kit.

ROSMAP

The Rush Religious Orders Study (ROS), started in 1994, enrolled Catholic priests, nuns, and brothers, from about 40 groups in 12 states3. Since January 1994, 1321 participants completed their baseline evaluation, of whom 1259 were non-Hispanic white. The follow-up rate of survivors exceeds 90%. Participants were free of known dementia at enrollment, agreed to annual clinical evaluations, and signed both an informed consent and

an Anatomic Gift Act form donating their brains at time of death 3. A more detailed description of ROS has been published previously3. Participants take a neuropsychological test battery. DNA was extracted from whole blood, lymphocytes, or frozen post-mortem brain tissue. Genotyping was performed at the Broad Institute's Center for Genotyping and the Translational Genomics Research Institute 2.

The Rush Memory and AP (MAP), started in 1997, enrolled older men and women from assisted living facilities in the Chicago area with no evidence on dementia at baseline 1. Since October 1997, 1815 participants completed their baseline evaluation, of whom 1701 were non-Hispanic white people. The follow-up rate of survivors exceeds 90%. Participants agreed to annual clinical evaluations, and signed both an informed consent and an Anatomic Gift Act form donating their brains at time of death. A more detailed description of the MAP has been published previously 1. Participants were invited to take a neuropsychological test battery. DNA was extracted from whole blood, lymphocytes, or frozen postmortem brain tissue. Genotyping was performed at the Broad Institute's Center for Genotyping and the Translational Genomics Research Institute 2.

SHIP

The Study of Health in Pomerania (SHIP) is a population-based project in West Pomerania, the north-east area of Germany[90,91]. A sample from the population aged 20 to 79 years was drawn from population registries. First, the three cities of the region (with 17,076 to 65,977 inhabitants) and the 12 towns (with 1,516 to 3,044 inhabitants) were selected, and then 17 out of 97 smaller towns (with less than 1,500 inhabitants), were drawn at random. Second, from each of the selected communities, subjects were drawn at random,

proportional to the population size of each community and stratified by age and gender. Only individuals with German citizenship and main residency in the study area were included. Finally, 7,008 subjects were sampled, with 292 persons of each gender in each of the twelve five-year age strata. In order to minimize drop-outs by migration or death, subjects were selected in two waves. The net sample (without migrated or deceased persons) comprised 6,267 eligible subjects. Selected persons received a maximum of three written invitations. In case of non-response, letters were followed by a phone call or by home visits if contact by phone was not possible. The SHIP population finally comprised 4,308 participants (corresponding to a final response of 68.8%).

UKB

The UK Biobank is a prospective cohort of over 500,000 individuals originating from 22 centers within the United Kingdom aged between 40 and 69[6]. Biological samples and physical measurements were collected and participants answered extensive questionnaires on health. Whole Exome Sequencing data was released in 2019 for 44,805 self-identified white individuals.

Cohort specific QC and GWAS

ARIC

Genotypes derived from the Affymetrix Genome-Wide Human SNP Array 6.0. Individuals were dropped if they refused DNA testing, > 5% missingness, or were identified as genetic outliers. Monomorphic SNPs, and SNPs with > 5% missingness were

dropped. Individuals with European ancestry were imputed to the Haplotype Reference Consortium while individuals of African ancestry were imputed to 1000G (March 2012).

GWAS was additionally adjusted for 10 principal components.

CHS

European ancestry Genotypes derived from the Illumina 370CNV chip, ITMAT-Broad-CARe Illumina iSELECT chip. Individuals were excluded from the sample due to the presence of baseline CHD, CHF, peripheral vascular disease, valvular heart disease, stroke or TIA, or lack of available DNA. After genotyping, individuals were excluded if they had a call rate <=95% or if their genotype was discordant with known sex or prior genotyping. SNPs were excluded if they had a call rate <97%, HWE P < 10-5, >2 duplicate errors or Mendelian inconsistencies, or heterozygote frequency = 0.

African ancestry genotypes were derived from the Illumina HumanOmni1-Quad_v1 BeadChip system. Individuals were excluded from the sample due to lack of available DNA. After genotyping, individuals were excluded if they had a call rate <=95% or if their genotype was discordant with known sex or prior genotyping. SNPs were excluded if they had a call rate <97%, HWE P < 10-5, >1 duplicate error or Mendelian inconsistency, or heterozygote frequency = 0.

African ancestry GWAS was additionally adjusted for 5 principal components.

FHS

Genotyping derived from the Affymetrix 500K mapping array plus Affymetrix 50K supplemental array. Genotyped SNPs removed based on: HWE p-value of less than

0.000001, call rate of less than 96.9%, MAF of less than 0.01, no mapping correctly from Build 36 to Build 37 locations, missing a physical location, Mendelian errors greater than 1000, not being on chromosomes 1-22 or X, duplicates. Imputed to 1000G (March 2012).

Due to family structure, a linear mixed effects model with fixed polygenic variance was used to account for family relationship during the GWAS.

MESA

Genotypes derived from the Affymetrix Genome-Wide Human SNP Array 6.0. Individuals with < 95% genotyping success rate were excluded while SNPs with < 90% genotyping success rate were excluded. Imputation was performed to 1000G (March 2012).

GWAS was additionally adjusted for 4 principal components.

ROSMAP

Genotypes derived from the Affymetrix Genome-Wide Human SNP Array 6.0. Sample-level quality control assessment included exclusion of samples with genotype success rate <95%, discordance between inferred and reported gender, and excess inter/intraheterozygosity. SNP-level quality control assessment included exclusion of SNPs with Hardy-Weighberg equilibrium ($p<0.001$), MAF < 0.01, genotype call rate < 0.95, misshap test < 1x10-9. Subsequently, EIGENSTRAT was used to identify and remove population outliers using default parameters

SHIP

Genotypes derived from the Affymetrix Genome-Wide Human SNP Array 6.0. Arrays with a call rate below 94%, duplicate samples as identified by estimated IBD as well as individuals with reported and genotyped gender mismatch were excluded. The final sample call rate was 99.51%.

GWAS was additionally adjusted for 10 PCs to account for population substructure.

UKB

Whole exome sequencing, sequencing alignment and variant identification was performed as previously described[92]. Genotypes were available through the Affymetrix UK Biobank Axiom array and the Affymetrix UK BiLEVE Axiom array. QC, phasing and imputation information have been described previously (http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580, http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=157020).

Due to its recruitment method and extremely large sample size, as much as 30% of the UKB dataset would need to be filtered due to typical relatedness or genetic ancestry filters[93]. To account for this, GWAS was performed using BOLT-LMM (v2.3.2) with the addition of a kinship matrix to control for related individuals[94].

Ethics Statements

ARIC

Institutional Review Board approvals were obtained by the coordinating center and each ARIC study center. The research was conducted in accordance with the principles

described in the Helsinki Declaration. All subjects in the ARIC study gave informed consent.

CHS

CHS was approved by institutional review committees at each field center and individuals in the present analysis had available DNA and gave informed consent including consent to use of genetic information for the study of cardiovascular disease.

FHS

The Boston University Medical Campus Institutional Review Board approved the FHS genome-wide genotyping.

MESA

MESA All MESA participants provided written and informed consent to participate in genetic studies. All study sites received approval to conduct this research from local Institutional Review Boards.

ROSMAP

All participants provided written informed consent and approval was obtained from the institutional review board.

SHIP

The study has been conducted according to the recommendations of the Declaration of Helsinki. The study protocol of SHIP was approved by the medical ethics committee of the University of Greifswald. Written informed consent was obtained from each of the study participants.

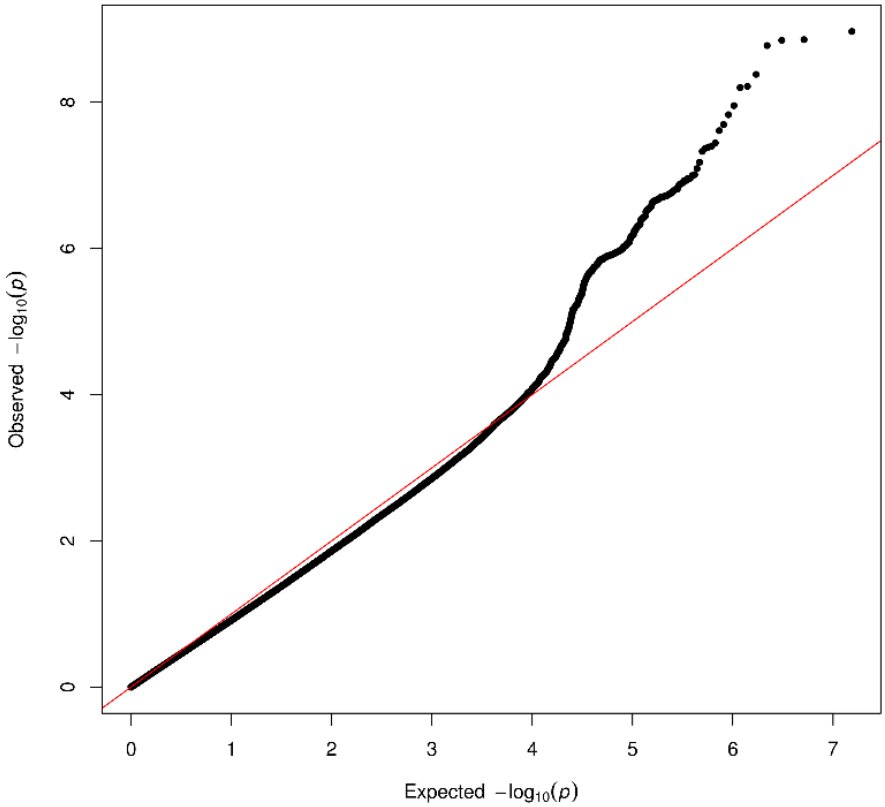
UKB

Research was approved by UK Biobank to ensure consistent with participant's consent and framework of data access.

Supplementary Table 3.1. UK Biobank cell counts

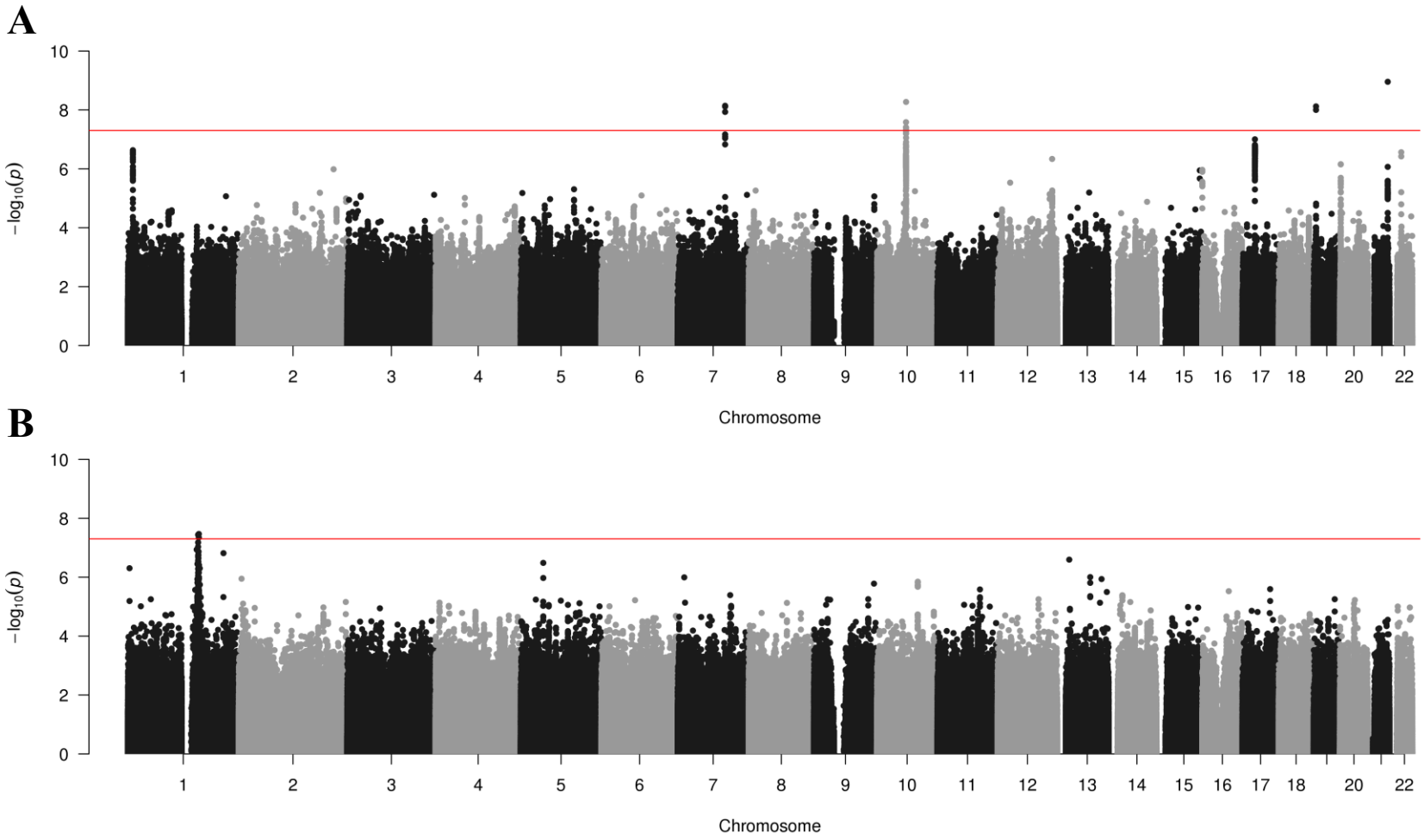
Participant Characteristics	UKBiobank
N	44,805
WBC (10^3/μl)	7.0 ± 1.7
Lymphocytes (10^3/μl)	2.0 ± 0.6
Monocytes (10^3/μl)	0.48 ± 0.16
Neutrophils (10^3/μl)	4.3 ± 1.4
Eosinophils (10^3/μl)	0.17 ± 0.12
Basophils (10^3/μl)	0.05 ± 0.04
Platelets (10^3/μl)	242.9 ± 55.4
RBC (10^6/μl)	4.5 ± 0.4
Nucleated RBC	247 (0.6)

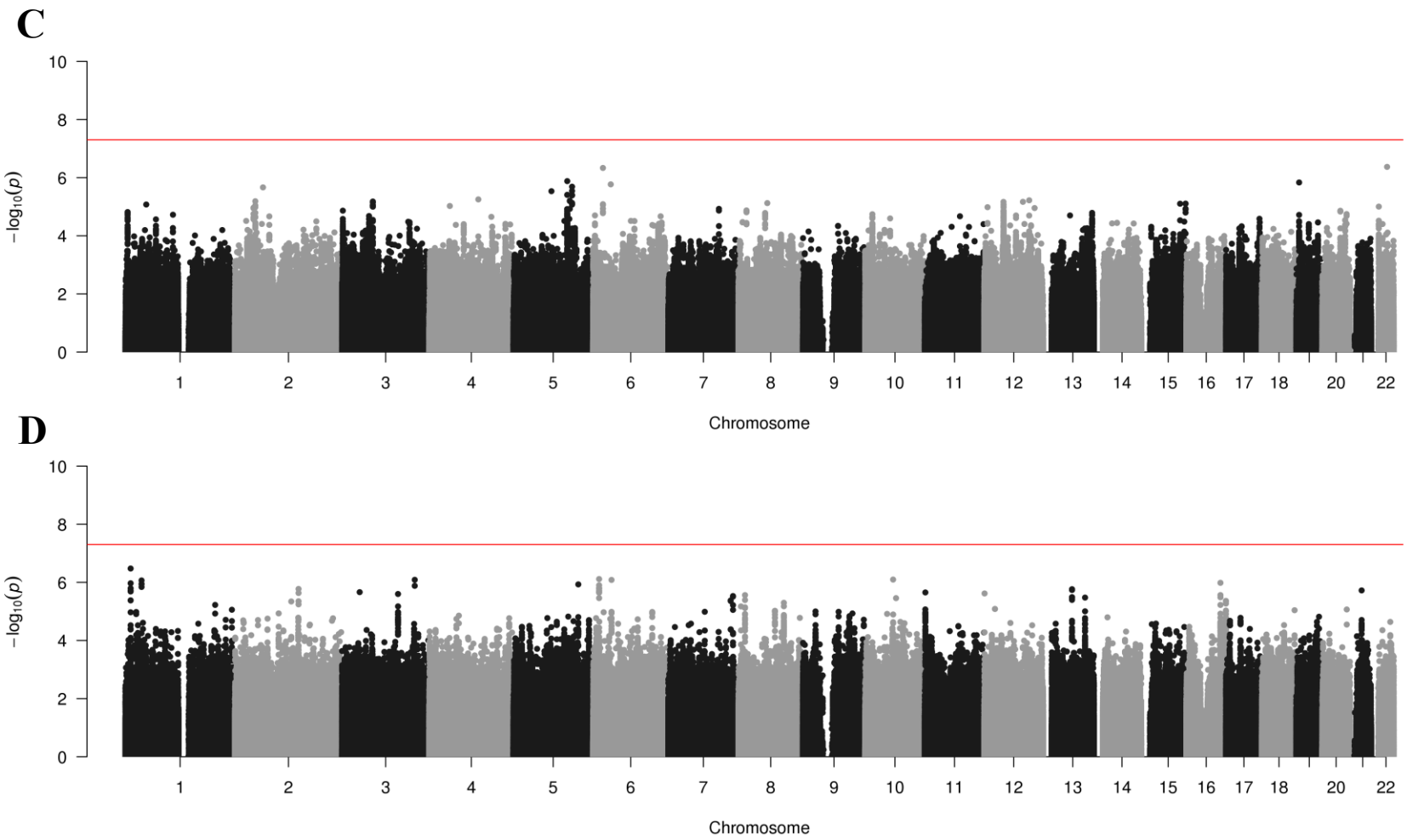
Values are number (%) or mean ± SD.
Abbreviations: WBC, white blood cell count; RBC, red blood cell count

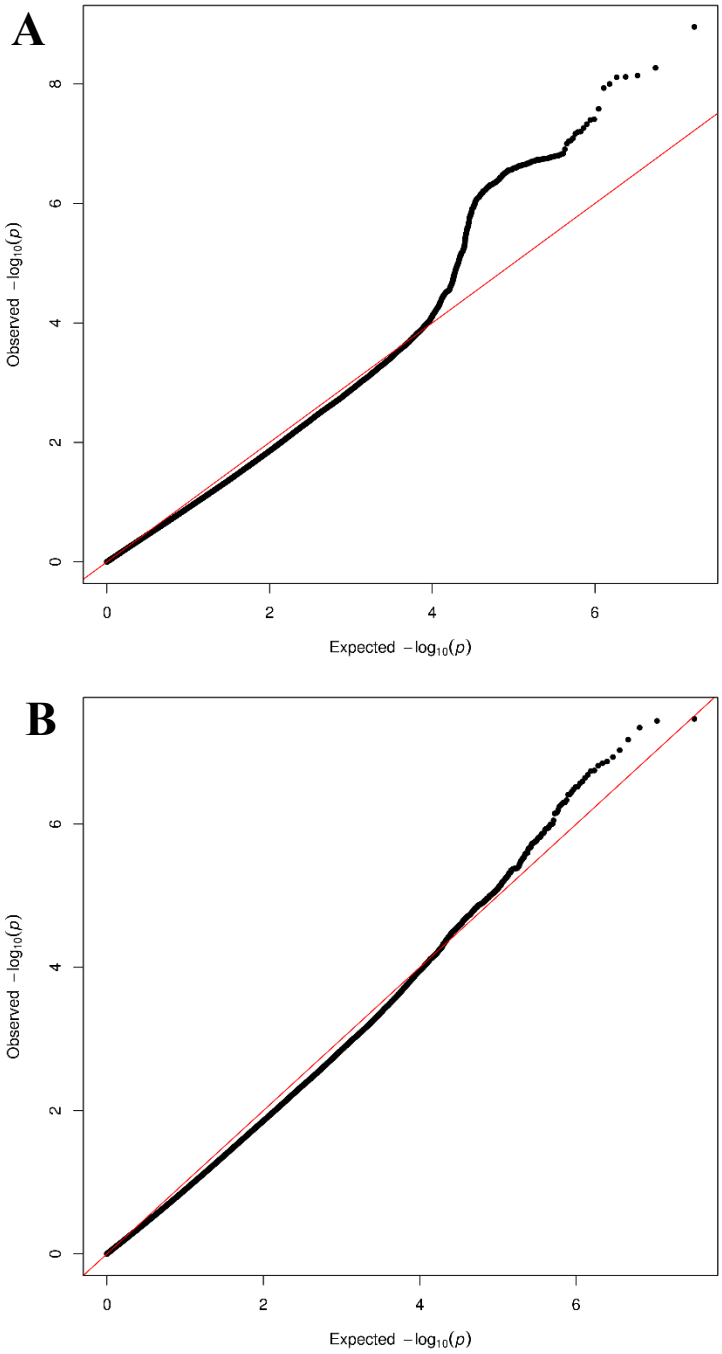


Supplementary Figure 3.1. QQ Plot of trans-ethnic meta-analysis. Highlights minimal inflation based on expected normal distribution

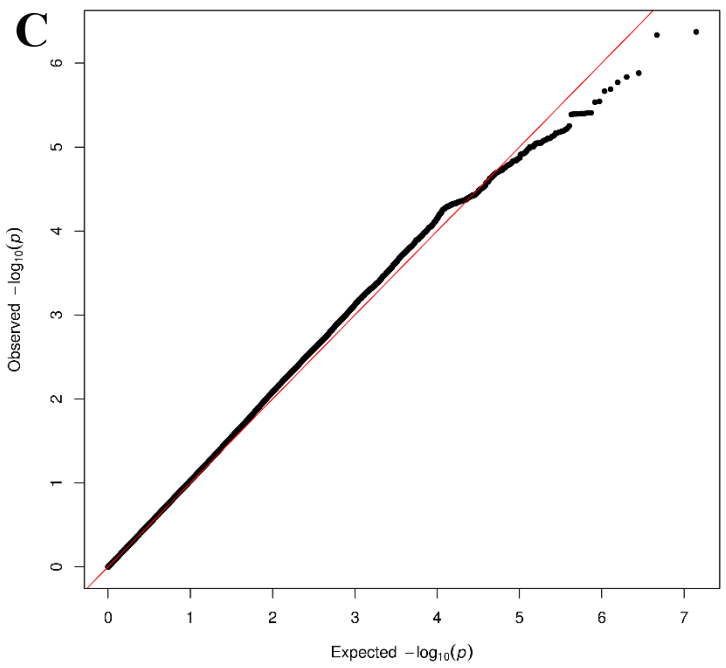
Supplementary Figure 4.2. Ethnicity-specific manhattan plots. Whites (A) and Blacks (B) were performed with a random effects meta-analysis to account for differences due to mtDNA-CN estimation platform. Chinese (C) and Hispanic (D) were present in MESA only. Filters: Imputation quality > 0.6, MAF > 0.005, SNP had to be present in at least half of the cohorts.

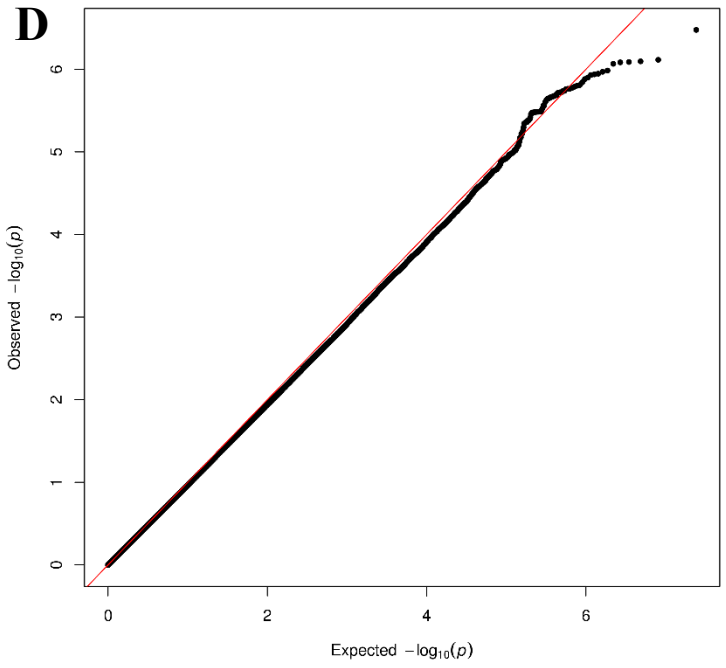

A
B
−log10(p)
Chromosome
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
22


C
D
10
8
6
4
2
0
−log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
22
Chromosome



Supplementary Figure 3.3. Ethnicity-specific QQ plots. Whites (A), Blacks (B), Chinese (C), and Hispanic (D)


C
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
1
2
3
4
5
6
7


D
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6

Chapter 4 :
Mitochondrial DNA heteroplasmy is associated with overall mortality and cardiovascular disease

INTRODUCTION

Mitochondria play a critical role in energy metabolism as the key organelle involved in oxidative phosphorylation. Declines in oxidative capacity represent a key feature of mitochondrial dysfunction which is known to play a critical role in the aging process and has been previously linked to several aging-related diseases[95–97]. As a major byproduct of oxidative phosphorylation, reactive oxygen species lead to mitochondrial DNA damage and heteroplasmy – the presence of multiple distinct mtDNA genomes within an individual. Although a proportion of heteroplasmy is inherited (~30%), the majority is acquired through somatic mutation (~70%)[98] and has been shown to increase in prevalence with age[99]. Furthermore, the frequency of a given heteroplasmy may rise and fall over the lifetime of an individual through stochastically biased mtDNA turnover[100].

Given that random mutations, on average, are detrimental due to a lack of selection, in addition to the absence of intergenic and intronic space within the mitochondrial genome, the accumulation of mitochondrial DNA mutations and heteroplasmy is likely to negatively impact mitochondrial function and contribute to mortality and disease. Indeed, several studies have shown mtDNA mutations lead to mitochondrial dysfunction and thus disease[80,101]. Additionally, mice with reduced DNA repair activity due to mutations in the *polG* editing domain experienced increased mortality and aging-related phenotypes such as premature balding, weight loss and grey hair[81].

Cardiovascular disease (CVD), an umbrella term including several pathologies relating to the vasculature, represents one of the leading causes of mortality and morbidity with an annual incidence of 800,000 within the United States[102]. Although CVD is

heterogenous, the primary pathological insult for most cases is the chronic inflammatory process known as atherosclerosis. Mitochondrial dysfunction, and more specifically mtDNA mutations, have been previously linked to the initiation, progression, and severity of atherosclerosis. Mitochondrial DNA damage has been shown to be associated with the degree of atherosclerosis in human aortas[103]. Additionally, *ApoE* knockout mouse models of hyperlipidemia developed mtDNA damage in the endothelial wall which preceded plaque formation[7]. Furthermore, bone marrow transplantation from *polG* knock out mice with elevated levels of mtDNA damage into *ApoE* knockout mice resulted in plaque instability, highlighting the potential mechanistic role of heteroplasmy in circulating cells[7].

In the present study, we leverage heteroplasmy calculated across 5,785 individuals from the Atherosclerosis Risk in Communities (ARIC) study to evaluate the role of heteroplasmic burden in mortality and CVD.

METHODS

Study populations

The ARIC study recruited 15,792 individuals aged 45 to 65 years from 4 US communities between 1987 and 1989. DNA for whole genome sequencing (WGS) was collected from different visits and was derived from buffy coat using the Gentra Puregene Blood Kit (Qiagen). Analyses were restricted to 6,659 individuals with whole genome sequence (WGS) data available from ARIC low-pass WGS efforts (n = 3,604) or WGS data sequenced through TOPMed (n = 3,055). We further excluded 874 individuals who

had missing phenotype information, potential contamination, insufficient coverage, or were duplicates for a final sample size of 5,785 (Supplementary Methods).

All participants provided written informed consent and all centers obtained approval from their respective institutional review boards.

Cardiovascular disease definition and adjudication

Event adjudication through 2017 in ARIC consisted of expert committee review of death certificates, hospital records and telephone interviews. Incident cardiovascular disease (CVD) was defined as either incident coronary artery disease (CAD) or incident stroke. Incident CAD was defined as first incident MI or death owing to CAD while incident stroke was defined as first nonfatal stroke or death due to stroke. Individuals in ARIC with prevalent CVD at baseline were excluded from incident analyses.

Whole genome sequencing

Whole genome sequencing (WGS) data available through TOPMed targeting a mean depth of at least 30x was generated using the Illumina HiSeq X Ten at six sequencing centers. Sequence reads were mapped to the GRCh38 reference genome using BWA[35]. Variant calling and quality control were performed as previously described[104].

Low pass WGS in ARIC generated at the Baylor College of Medicine Human Genome Sequencing Center using Nano or PCR-free DNA libraries on the Illumina HiSeq 2000 resulted in 6x average sequencing depth. Sequence reads were mapped to the hg19 reference genome using BWA[35]. Variant calling and quality control were performed as previously described[40].

To ensure sufficient coverage across the entire genome, reads were remapped to a circularized version of the mitochondrial genome. Potential contamination by nuclear mitochondrial DNA segments (NUMTs) was limited by removing any mitochondrial reads which additionally aligned to the nuclear genome[105]. GATK indel realigner was run to minimize misaligned reads and Picard deduplication was used to remove duplicate reads[39,106].

Statistical analyses

Cox-proportional hazards regression was used to estimate hazard ratios (HRs) for mortality and incident disease outcomes. Follow-up time was defined from DNA collection date through loss to follow-up, death, or study end point (2017 in ARIC). All statistical analyses were performed using R (version 3.3.3).

RESULTS

Sample characteristics

After quality control, the study included 5,785 participants from ARIC (Table 1). The mean age of study participants was 58 years, 56.2% of participants were women (n = 3,249) and 32.8% were black (n = 1,897).

Distribution of heteroplasmy

The distribution of heteroplasmy was highly skewed with 69.3% of individuals having no detectable heteroplasmy at the 5% allele frequency level (n = 4,011) (**Supplementary Table 4.1**). Across all individuals, heteroplasmy was evenly distributed

across the mitochondrial genome except for the D-loop region (**Figure 4.1**). This observation was expected as the D-loop is known to be a highly polymorphic hypervariable region.

Mortality

A total of 2,989 deaths were observed in ARIC during 115,626 person-years of follow-up. In an age, sex, ethnicity, sequencing depth and DNA collection site adjusted cox proportional-hazards inverse weighted meta-analysis of results from both ARIC matches we observed a statistically significant association when comparing mortality in individuals with and without a heteroplasmic site (HR: 1.19; 95% CI: 1.10-1.28; *p* value: 1.09×10^{-5}). A more stringent model was applied controlling for traditional risk factors including BMI, systolic blood pressure, smoking status, LDL, HDL, triglycerides, prevalent diabetes, use of hypertension medication, and history of myocardial infarction, resulted in a HR of 1.17 (95% CI: 1.09-1.27; *p* value: 5.5×10^{-5}) (**Figure 4.2**). Importantly, these results are not driven by a single batch as we observe significant HRs of 1.21 and 1.14 for both the low pass and TOPMed batches in the more stringent model (95% CI: 1.08-1.35, 1.02-1.26; *p* value: 6.7×10^{-4}, 0.02) (**Supplementary Figure 4.1**).

Heteroplasmic mutational burden

To explore the impact of heteroplasmic mutational burden on mortality we explored the role of heteroplasmies predicted to be synonymous, nonsynonymous neutral and nonsynonymous deleterious. Of the 1,774 heteroplasmies we identified, 160 were predicted to be pathogenic as defined by a scaled Combined Annotation Dependent

Depletion (CADD) score > 15. In a meta-analysis of both ARIC batches adjusting for all demographics and traditional risk factors we observe a HR of 1.26 (95% CI: 1.07-1.47; *p* value: 4.3×10^{-3}) when comparing mortality in individuals with at least one deleterious heteroplasmy compared to those without any heteroplasmies (Supplemental Figure 2). A meta-analysis in individuals with synonymous heteroplasmies observed a HR of 1.15 (95% CI: 1.05-1.25; *p* value: 1.8×10^{-3}) (**Figure 4.3**).

To further address heteroplasmic mutational burden we investigated the impact of having multiple heteroplasmies. A total of 358 individuals were observed to have more than one heteroplasmy between both ARIC batches. A meta-analysis adjusting for all covariates within ARIC revealed individuals with only one heteroplasmy have a HR of 1.14 (95% CI: 1.05-1.24; *p* value: 1.6×10^{-3}) compared to individuals with no heteroplasmies, while individuals with 2 or more heteroplasmies have a HR of 1.28 (95% CI: 1.12-1.47; *p* value: 4.5×10^{-4}) (**Figure 4.4**). Although these two values are not significantly different (*p* value = 0.15), the directionality of this effect suggests an additive effect for heteroplasmic burden.

Incident cardiovascular disease

Within ARIC 1,097 incident CVD events were observed (525 and 572 for ARIC low pass and ARIC TOPMed respectively). In an inverse-weighted meta-analysis exploring incident CVD events adjusting for age, sex, ethnicity, sequencing depth and DNA collection site we observed a HR of 1.07 (95% CI: 0.94-1.21, *p* value: 0.29) when comparing individuals with at least one heteroplasmic site compared to individuals without a heteroplasmy. However, once we adjusted for traditional risk factors including BMI, systolic blood pressure, smoking status, LDL, HDL, triglycerides, prevalent diabetes and

use of hypertension medication we observed a significant HR of 1.14 (95% CI: 1.00-1.29, *p* value: 0.048) (**Figure 4.5**). The change in significance from the base model to the full model most likely highlights the complex nature of mitochondrial dysfunction where there is not only a relationship with CVD risk, but also with several traditional risk factors such as lipid levels[107] or BMI[108].

We further investigated CVD by exploring the impact of heteroplasmy on its component parts – coronary artery disease (CAD) and stroke. While we observed no significant relationship between heteroplasmy and incident CAD (n = 758) in a meta-analysis adjusting for traditional risk factors and demographics (HR: 1.06; 95% CI: 0.91-1.24; *p* value: 0.46), we did observe a relationship with incident stroke (n = 538) with a HR of 1.37 (95% CI: 1.15-1.64; *p* value: 4.6×10^{-4}) (**Supplementary Figure 4.3**).

DISCUSSION

We explored the impact of heteroplasmic burden measured on mortality and incident CVD in 5,785 self-identified white and black individuals from the ARIC study. Individuals with at least one heteroplasmy were 17% more likely to die during follow up and 14% more likely to develop CVD compared to individuals without any heteroplasmy independent of traditional risk factors. Furthermore, heteroplasmy was associated with mortality independent of CVD deaths (**Supplementary Figure 4.4**), potentially highlighting its importance in other major contributors to mortality such as cancer and respiratory disease. While previous research has shown the impact of specific heteroplasmies on mortality[20], to our knowledge this is the first time overall heteroplasmic burden has been shown to be associated with mortality and disease.

Our analyses on heteroplasmic mutational burden highlight the predicted pathogenicity of a heteroplasmy has a large impact on patient outcome. Counterintuitively, synonymous heteroplasmies were also associated with mortality, however research has shown silent mutations may modulate transcription and translation[109,110]. The observed impact of synonymous heteroplasmies on mortality could also reflect a hypermutable state which may impact nuclear genes. Although event rates were too low to accurately explore heteroplasmic mutational burden in CVD, future research should explore this potential relationship to further understand the role of heteroplasmy in cardiac disease.

Although stroke and CAD have overlapping risk factors, our analyses found while the presence of heteroplasmy was significantly associated with incident stroke the same was not true for CAD. This finding may be driven by the more diverse etiology of stroke compared to CAD[111]. While CAD is mainly driven within the context of coronary atherosclerosis, stroke occurs within the setting of atherosclerosis, small vessel disease and cardiac embolism. Our results may highlight that heteroplasmy is associated with non-atherosclerotic cardiovascular disease.

In conclusion, our study demonstrates heteroplasmy is associated with mortality and incident CVD independent of traditional risk factors. Furthermore, our observed associations also occur independently of mtDNA-CN, suggesting our mtDNA quantity and quality capture different facets of mitochondrial dysfunction (**Supplementary Figure 4.5**). We also observed the effect of heteroplasmy on mortality to be additive and elevated risk is present regardless of the predicted pathogenicity of the heteroplasmic site. While this study was not designed to address the mechanism of action or a potential causal

relationship between heteroplasmy and disease, our results are in line with previous research highlighting the role of mitochondrial dysfunction in mortality and CVD.

FIGURES AND TABLES

Table 4.1. Participant characteristics

Participant Characteristics	ARIC low pass	ARIC TOPMed
N	2,984	2,801
Age	57.5 ± 6.0	58.7 ± 5.9
Sex (female)	1,791 (60.0)	1,458 (52.1)
Race (black)	1,707 (57.2)	190 (6.8)
BMI, kg/m2	29.1 ± 6.0	27.6 ± 5.1
Systolic BP	124.9 ± 20.1	122.5 ± 18.5
LDL	133.3 ± 36.4	134.8 ± 37.3
HDL	52.1 ± 16.9	48.5 ± 16.3
TG	121.3 ± 62.5	136.4 ± 65.5
Current smoker	691 (23.2)	692 (24.7)
Prevalent diabetes	546 (18.3)	405 (14.5)
History of MI	170 (5.7)	221 (7.9)
Hypertension Medication	1,175 (39.4)	958 (34.2)

Values are number (%) or mean ± SD
Abbreviations: SD, standard deviation; BMI, body mass index; BP, blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TG, triglycerides; MI, myocardial infarction

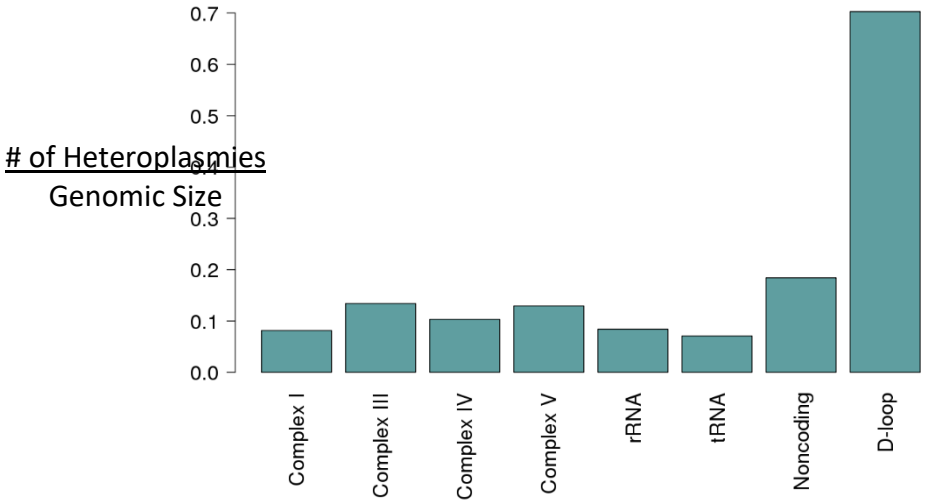


Figure 4.1. Distribution of heteroplasmy across mitochondrial genome. Heteroplasmy is approximately evenly distributed across the mitochondrial genome with regards to genomic size with exception to the hypervariable D-loop.

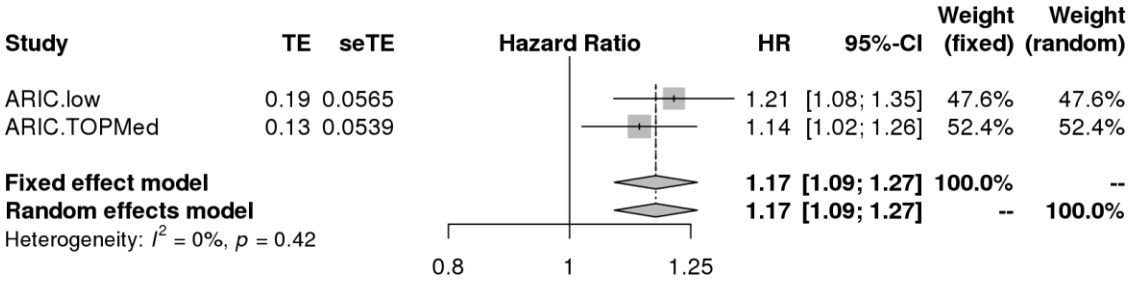


Figure 4.2. Effect of heteroplasmy on mortality. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for individuals with at least one heteroplasmy compared to individuals without any heteroplasmy.

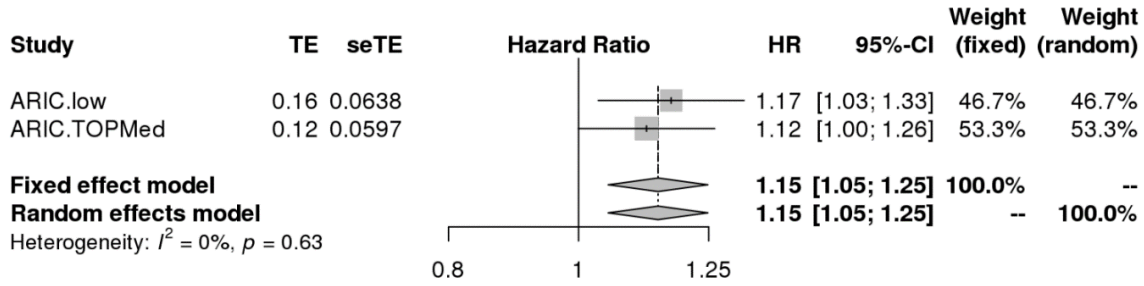


Figure 4.3. Effect of synonymous heteroplasmies on mortality. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for synonymous heteroplasmies compared to individuals without heteroplasmy.

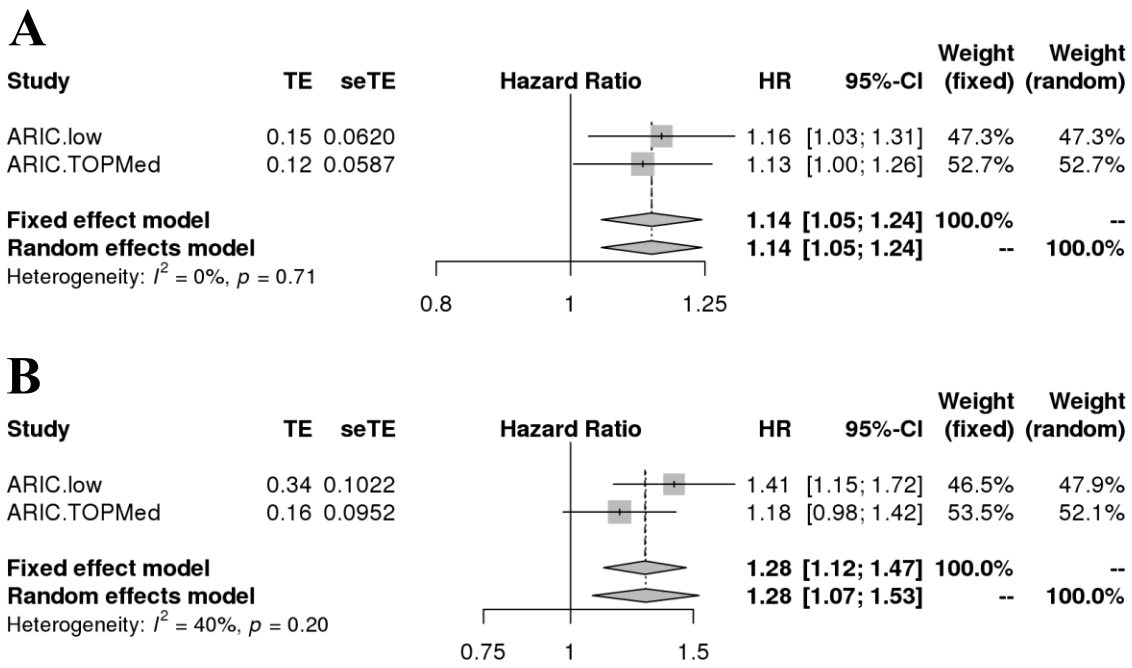


Figure 4.4. Effect of multiple heteroplasmies on mortality. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for a single heteroplasmy (A) and two or more heteroplasmies (B) compared to individuals without heteroplasmy.

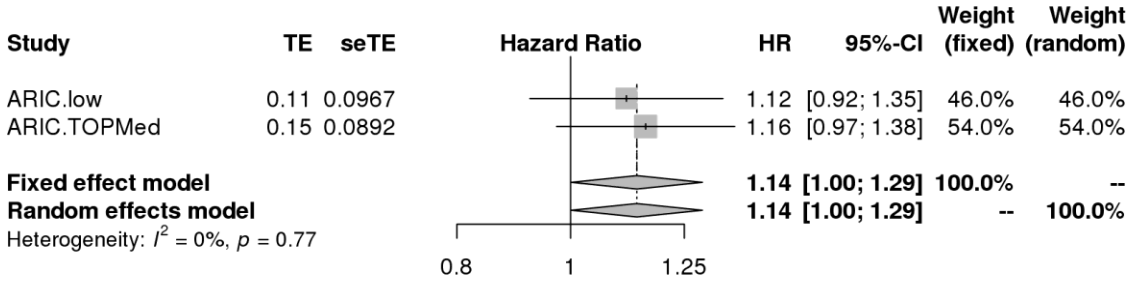


Figure 4.5. Effect of multiple heteroplasmies on incident cardiovascular disease. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for individuals with at least one heteroplasmy compared to individuals without any heteroplasmy.

SUPPLEMENTARY MATERIAL

Supplementary Methods

Low level contamination caused by sample mix-up, cross-contamination, carry-over during WGS runs, or postprocessing issues during adapter removal could influence heteroplasmy calls. To avoid misinterpretation, mtDNA-server implements an intra-sample contamination check to identify potentially contaminated samples. Two haplogroup profiles are generated; a minor profile based on heteroplasmy calls < 50% allele fraction and a major profile based on heteroplasmy calls > 50% allele fraction. Homoplasmic variants are included to augment the haplogroup calls. Individuals with different haplogroups between minor and major profiles are flagged as potentially contaminated by the mtDNA-server pipeline. Flagged samples were manually curated by evaluating the phylotree distance between the minor and major profiles, the number of heteroplasmies called in a sample, and the average heteroplasmic frequency.

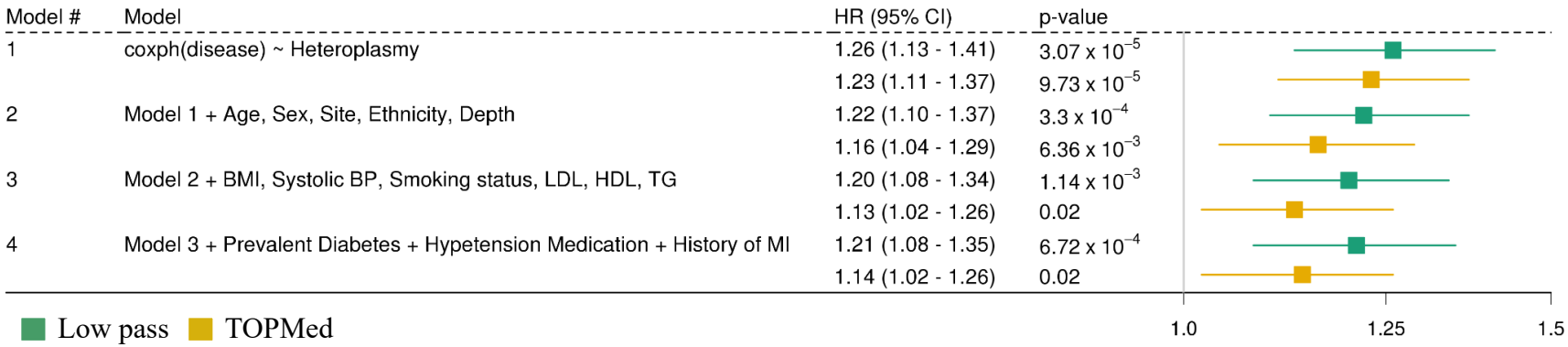
As additionally methods to identify potential contamination, samples were filtered if they contained a large proportion of heteroplasmies which have never been observed in phylotree, several heteroplasmies associated with a different haplogroup then the one identified for the sample or if a majority of the variants used to call the haplogroup for a sample are heteroplasmies.

Of the 6,659 WGS samples available in ARIC, 223 were filtered due to haplogroup contamination, 101 were removed due to poor haplogroup quality, and 473 were removed due to missing phenotype information. Additionally, 77 samples which were

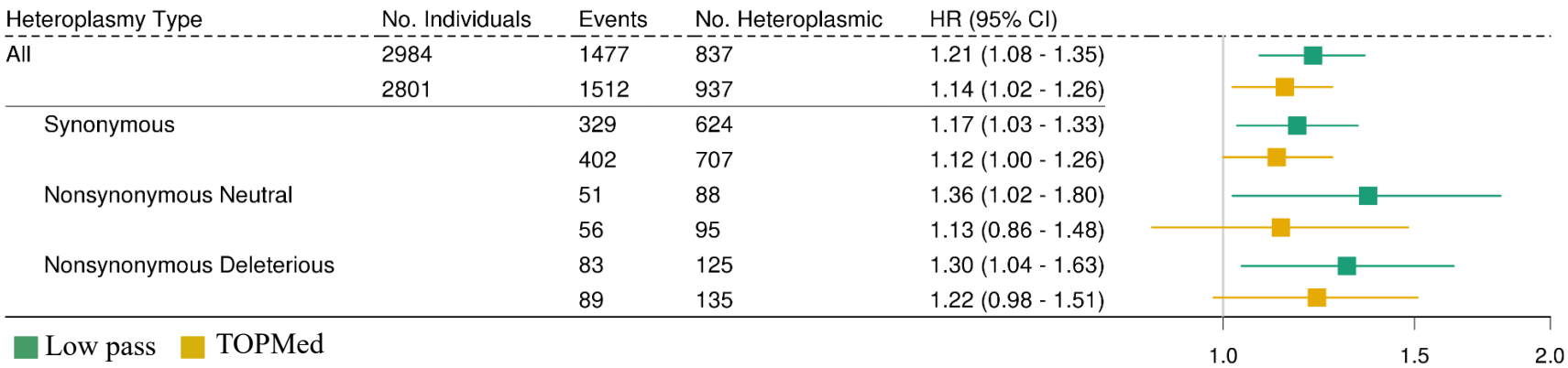
present in both the low pass data and the TOPMed datasets were removed from the low pass batch. Our final sample size for the low pass and TOPMed batches were 2,984 and 2,801 respectively (N = 5,785).

Supplemental Table 4.1. Population distribution of heteroplasmy

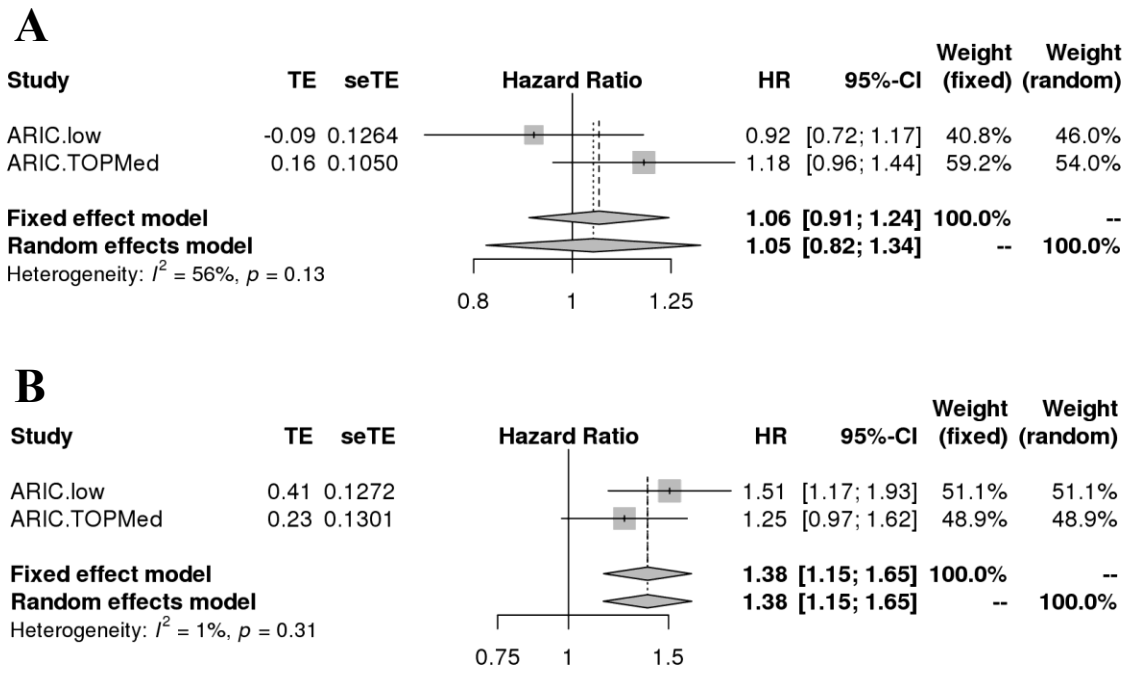
Heteroplasmies	Individuals (low pass)	Individuals (TOPMed)
0	2,147	1,864
1	678	738
2	131	167
3	22	26
4	3	4
5	2	2
6	1	0



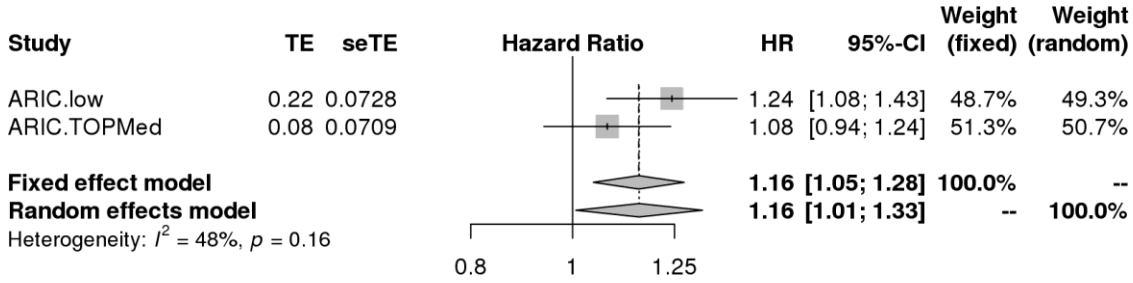
Supplemental Figure 4.1. Effect of having single heteroplasmy on overall mortality. Hierarchical cox proportional-hazards model building of traditional risk factors for effect of heteroplasmy on mortality. Hazard ratio (HR) and 95% confidence interval (CI) represented on the far right graphically. Abbreviations: BMI, body mass index; BP, blood pressure; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TG, triglycerides; MI, myocardial infarction



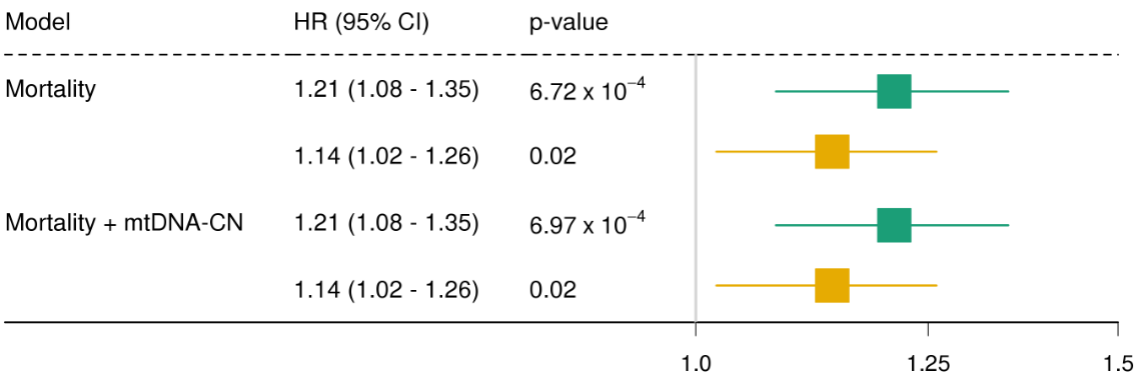
Supplemental Figure 4.2. Effect of having single heteroplasmy on overall mortality broken down by predicted mutational burden. Hierarchical cox proportional-hazards model adjusting for age, sex, ethnicity, sequencing depth, body mass index, systolic blood pressure, smoking status, low-density lipoprotein, high-density lipoprotein, triglycerides, prevalent diabetes, hypertension medication status and history of myocardial infarction. Hazard ratio (HR) and 95% confidence interval (CI) represented on the far right graphically.



Supplemental Figure 4.3. Effect of heteroplasmy on incident CAD and stroke. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for individuals with at least one heteroplasmy compared to individuals without any heteroplasmy for CAD (A) and Stroke (B).



Supplemental Figure 4.4. Effect of heteroplasmy on non-CVD mortality. Inverse-variance weighted meta-analysis of ARIC low pass sequencing and ARIC TOPMed sequencing batches for individuals with at least one heteroplasmy compared to individuals without any heteroplasmy.



Supplemental Figure 4.5. mtDNA-CN does not affect impact of heteroplasmy on mortality. Hierarchical cox proportional-hazards model adjusting for age, sex, ethnicity, sequencing depth, body mass index, systolic blood pressure, smoking status, low-density lipoprotein, high-density lipoprotein, triglycerides, prevalent diabetes, hypertension medication status and history of myocardial infarction. Hazard ratio (HR) and 95% confidence interval (CI) represented on the far right graphically.

References

BIBLIOGRAPHY

1. Harman, D. The biologic clock: the mitochondria? *J. Am. Geriatr. Soc.* **20**, 145–147 (1972).
2. Honjo, I., Ozawa, K., Kitamura, O., Sakai, A. & Ohsawa, T. Rapid change of phospholipid in pancreas mitochondria during aging. *J. Biochem. (Tokyo)* **64**, 311–320 (1968).
3. Ernster, L., Low, H., Nordenbrand, K. & Ernster, B. A component promoting oxidative phosphorylation, released from mitochondria during aging. *Exp. Cell Res.* **9**, 348–349 (1955).
4. Gómez-Serrano, M., Camafeita, E., Loureiro, M. & Peral, B. Mitoproteomics: Tackling Mitochondrial Dysfunction in Human Disease. *Oxid. Med. Cell. Longev.* **2018**, (2018).
5. Yu, E., Mercer, J. & Bennett, M. Mitochondria in vascular disease. *Cardiovasc. Res.* **95**, 173–182 (2012).
6. Zharikov, S. & Shiva, S. Platelet mitochondrial function: from regulation of thrombosis to biomarker of disease. *Biochem. Soc. Trans.* **41**, 118–123 (2013).
7. Yu, E. *et al.* Mitochondrial DNA damage can promote atherosclerosis independently of reactive oxygen species through effects on smooth muscle cells

and monocytes and correlates with higher-risk plaques in humans. *Circulation* **128**, 702–712 (2013).

8. Schleicher, M. *et al.* Prohibitin-1 maintains the angiogenic capacity of endothelial cells by regulating mitochondrial function and senescence. *J. Cell Biol.* **180**, 101–112 (2008).
9. Kurz, D. J. *et al.* Chronic oxidative stress compromises telomere integrity and accelerates the onset of senescence in human endothelial cells. *J. Cell Sci.* **117**, 2417–2426 (2004).
10. Erusalimsky, J. D. Vascular endothelial senescence: from mechanisms to pathophysiology. *J. Appl. Physiol. Bethesda Md 1985* **106**, 326–332 (2009).
11. Malik, A. N. & Czajka, A. Is mitochondrial DNA content a potential biomarker of mitochondrial dysfunction? *Mitochondrion* **13**, 481–492 (2013).
12. Van Houten, B., Hunter, S. E. & Meyer, J. N. Mitochondrial DNA damage induced autophagy, cell death, and disease. *Front. Biosci. Landmark Ed.* **21**, 42–54 (2016).
13. Guha, M. & Avadhani, N. G. Mitochondrial retrograde signaling at the crossroads of tumor bioenergetics, genetics and epigenetics. *Mitochondrion* **13**, 577–591 (2013).
14. Ashar, F. N. *et al.* Association of Mitochondrial DNA Copy Number With Cardiovascular Disease. *JAMA Cardiol.* **2**, 1247–1255 (2017).
15. Clay Montier, L. L., Deng, J. J. & Bai, Y. Number matters: control of mammalian mitochondrial DNA copy number. *J. Genet. Genomics Yi Chuan Xue Bao* **36**, 125–131 (2009).

16. Wei, Y.-H., Lu, C.-Y., Lee, H.-C., Pang, C.-Y. & Ma, Y.-S. Oxidative Damage and Mutation to Mitochondrial DNA and Age-dependent Decline of Mitochondrial Respiratory Functiona. *Ann. N. Y. Acad. Sci.* **854**, 155–170 (1998).

17. Richter, C. Oxidative damage to mitochondrial DNA and its relationship to ageing. *Int. J. Biochem. Cell Biol.* **27**, 647–653 (1995).

18. Mecocci, P. *et al.* Oxidative damage to mitochondrial DNA shows marked age-dependent increases in human brain. *Ann. Neurol.* **34**, 609–616 (1993).

19. Yakes, F. M. & Van Houten, B. Mitochondrial DNA damage is more extensive and persists longer than nuclear DNA damage in human cells following oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 514–519 (1997).

20. Tranah, G. J. *et al.* Mitochondrial DNA m.3243A > G heteroplasmy affects multiple aging phenotypes and risk of mortality. *Sci. Rep.* **8**, (2018).

21. Mancuso, M. *et al.* Mitochondrial m.3243A > G mutation and carotid artery dissection. *Mol. Genet. Metab. Rep.* **9**, 12–14 (2016).

22. Chen, S. *et al.* Association between leukocyte mitochondrial DNA content and risk of coronary heart disease: A case-control study. *Atherosclerosis* **237**, 220–226 (2014).

23. Pyle, A. *et al.* Reduced mitochondrial DNA copy number is a biomarker of Parkinson's disease. *Neurobiol. Aging* **38**, 216.e7-216.e10 (2016).

24. Wei, W. *et al.* Mitochondrial DNA point mutations and relative copy number in 1363 disease and control human brains. *Acta Neuropathol. Commun.* **5**, (2017).

25. Reznik, E. *et al.* Mitochondrial DNA copy number variation across human cancers. *eLife* **5**,

26. Hertweck, K. L. & Dasgupta, S. The Landscape of mtDNA Modifications in Cancer: A Tale of Two Cities. *Front. Oncol.* **7**, 262 (2017).

27. Thyagarajan, B., Wang, R., Barcelo, H., Koh, W.-P. & Yuan, J.-M. Mitochondrial copy number is associated with colorectal cancer risk. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **21**, 1574–1581 (2012).

28. Knez, J. *et al.* Correlates of Peripheral Blood Mitochondrial DNA Content in a General Population. *Am. J. Epidemiol.* **183**, 138–146 (2016).

29. Tin, A. *et al.* Association between Mitochondrial DNA Copy Number in Peripheral Blood and Incident CKD in the Atherosclerosis Risk in Communities Study. *J. Am. Soc. Nephrol. JASN* **27**, 2467–2473 (2016).

30. Ashar, F. N. *et al.* Association of mitochondrial DNA levels with frailty and all-cause mortality. *J. Mol. Med. Berl. Ger.* **93**, 177–186 (2015).

31. Google Scholar. Search Terms: 'mitochondrial DNA copy number', 'mitochondrial DNA content'. Available at: https://scholar.google.com. (Accessed: 21st February 2019)

32. Cai, N. *et al.* Genetic Control over mtDNA and Its Relationship to Major Depressive Disorder. *Curr. Biol.* **25**, 3170–3177 (2015).

33. Guo, W., Jiang, L., Bhasin, S., Khan, S. M. & Swerdlow, R. H. DNA Extraction Procedures Meaningfully Influence qPCR-Based mtDNA Copy Number Determination. *Mitochondrion* **9**, 261–265 (2009).

34. MitoPipeline: Generating Mitochondrial copy number estimates from SNP array data in Genvisis. Available at: http://genvisis.org/MitoPipeline/. (Accessed: 27th November 2017)

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

36. Yu, B. *et al.* Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine Levels and Incident Coronary Heart Disease. *Circ. Cardiovasc. Genet.* **8**, 351–355 (2015).

37. Ding, J. *et al.* Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLOS Genet.* **11**, e1005306 (2015).

38. Qian, Y. *et al.* fastMitoCalc: an ultra-fast program to estimate mitochondrial DNA copy number from whole-genome sequences. *Bioinformatics* **33**, 1399–1401 (2017).

39. Picard Tools - By Broad Institute. Available at: https://broadinstitute.github.io/picard/. (Accessed: 4th January 2019)

40. Morrison, A. C. *et al.* Whole Genome Sequence-Based Analysis of a Model Complex Trait, High Density Lipoprotein Cholesterol. *Nat. Genet.* **45**, 899–901 (2013).

41. Duncan Temple Lang and the CRAN team (2018). RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.95-4.11. https://CRAN.R-project.org/package=RCurl.

42. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

43. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

44. Chanda, P., Huang, H., Arking, D. E. & Bader, J. S. Fast Association Tests for Genes with FAST. *PLOS ONE* **8**, e68585 (2013).

45. Gouhier, T. C. & Guichard, F. Synchrony: quantifying variability in space and time. *Methods Ecol. Evol.* **5**, 524–533 (2014).

46. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Res.* **11**, 1095–1099 (2001).

47. Zhang, Y. *et al.* Associations of mitochondrial haplogroups and mitochondrial DNA copy numbers with end-stage renal disease in a Han population. *Mitochondrial DNA Part A* **28**, 725–731 (2017).

48. Lee, H. K. *et al.* Decreased mitochondrial DNA content in peripheral blood precedes the development of non-insulin-dependent diabetes mellitus. *Diabetes Res. Clin. Pract.* **42**, 161–167 (1998).

49. Sookoian, S. *et al.* Epigenetic regulation of insulin resistance in nonalcoholic fatty liver disease: Impact of liver methylation of the peroxisome proliferator–activated receptor γ coactivator 1α promoter. *Hepatology* **52**, 1992–2000 (2010).

50. Memon, A. A. *et al.* Quantification of mitochondrial DNA copy number in suspected cancer patients by a well optimized ddPCR method. *Biomol. Detect. Quantif.* **13**, 32–39 (2017).

51. Ye, W. *et al.* Accurate quantitation of circulating cell-free mitochondrial DNA in plasma by droplet digital PCR. *Anal. Bioanal. Chem.* **409**, 2727–2735 (2017).

52. Hurtado-Roca, Y. *et al.* Adjusting MtDNA Quantification in Whole Blood for Peripheral Blood Platelet and Leukocyte Counts. *PLOS ONE* **11**, e0163770 (2016).

53. Nacheva, E. *et al.* DNA isolation protocol effects on nuclear DNA analysis by microarrays, droplet digital PCR, and whole genome sequencing, and on mitochondrial DNA copy number estimation. *PLoS ONE* **12**, (2017).

54. Tin, A. *et al.* Association between Mitochondrial DNA Copy Number in Peripheral Blood and Incident CKD in the Atherosclerosis Risk in Communities Study. *J. Am. Soc. Nephrol. JASN* **27**, 2467–2473 (2016).

55. Tang, Y. *et al.* Rearrangements of Human Mitochondrial DNA (mtDNA): New Insights into the Regulation of mtDNA Copy Number and Gene Expression. *Mol. Biol. Cell* **11**, 1471–1485 (2000).

56. Carling, P. J., Cree, L. M. & Chinnery, P. F. The implications of mitochondrial DNA copy number regulation during embryogenesis. *Mitochondrion* **11**, 686–692 (2011).

57. Harvey, A., Gibson, T., Lonergan, T. & Brenner, C. Dynamic regulation of mitochondrial function in preimplantation embryos and embryonic stem cells. *Mitochondrion* **11**, 829–838 (2011).

58. Copeland, W. C. Defects of Mitochondrial DNA Replication. *J. Child Neurol.* **29**, 1216–1224 (2014).

59. Alvarez, V. *et al.* Mitochondrial transcription factor A (TFAM) gene variation in Parkinson's disease. *Neurosci. Lett.* **432**, 79–82 (2008).

60. Mandel, H. *et al.* The deoxyguanosine kinase gene is mutated in individuals with depleted hepatocerebral mitochondrial DNA. *Nat. Genet.* **29**, 337–341 (2001).

61. Wang, L. *et al.* Molecular insight into mitochondrial DNA depletion syndrome in two patients with novel mutations in the deoxyguanosine kinase and thymidine kinase 2 genes. *Mol. Genet. Metab.* **84**, 75–82 (2005).

62. Workalemahu, T. *et al.* Genetic Variations Related to Maternal Whole Blood Mitochondrial DNA Copy Number: A Genome-Wide and Candidate Gene Study. *J. Matern.-Fetal Neonatal Med. Off. J. Eur. Assoc. Perinat. Med. Fed. Asia Ocean. Perinat. Soc. Int. Soc. Perinat. Obstet.* **30**, 2433–2439 (2017).

63. Guyatt, A. L. *et al.* A genome-wide association study of mitochondrial DNA copy number in two population-based cohorts. *Hum. Genomics* **13**, 6 (2019).

64. Longchamps, R. J. *et al.* Evaluation of mitochondrial DNA copy number estimation techniques. *bioRxiv* 610238 (2019). doi:10.1101/610238

65. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).

66. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

67. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.* **18**, 86–91 (2015).

68. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

69. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

70. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

71. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

72. de Coo, R. F. M., Buddiger, P., Smeets, H. J. M. & van Oost, B. A. Molecular Cloning and Characterization of the Human Mitochondrial NADH:Oxidoreductase 10-kDa Gene (NDUFV3). *Genomics* **45**, 434–437 (1997).

73. Bis, J. C. *et al.* Sequencing of 2 subclinical atherosclerosis candidate regions in 3669 individuals: Cohorts for Heart and Aging Research in Genomic Epidemiology

(CHARGE) Consortium Targeted Sequencing Study. *Circ. Cardiovasc. Genet.* **7**, 359–364 (2014).

74. Wilson, P. D., Franks, L. M., Cottell, D. C. & Benham, F. Alkaline phosphatase in mitochondria. *Cell Biol. Int. Rep.* **1**, 85–92 (1977).

75. Pirola, C. J. *et al.* Epigenetic Modifications in the Biology of Nonalcoholic Fatty Liver Disease: The Role of DNA Hydroxymethylation and TET Proteins. *Medicine (Baltimore)* **94**, e1480 (2015).

76. Tiao, M.-M. *et al.* Early stage of biliary atresia is associated with significant changes in 8-hydroxydeoxyguanosine and mitochondrial copy number. *J. Pediatr. Gastroenterol. Nutr.* **45**, 329–334 (2007).

77. Song, S. *et al.* DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4990–4995 (2005).

78. Wheeler, L. J. & Mathews, C. K. Nucleoside Triphosphate Pool Asymmetry in Mammalian Mitochondria. *J. Biol. Chem.* **286**, 16992–16996 (2011).

79. Mathews, C. K. DNA precursor metabolism and genomic stability. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **20**, 1300–1314 (2006).

80. Sevini, F. *et al.* mtDNA mutations in human aging and longevity: Controversies and new perspectives opened by high-throughput technologies. *Exp. Gerontol.* **56**, 234–244 (2014).

81. Kujoth, G. C. *et al.* Mitochondrial DNA Mutations, Oxidative Stress, and Apoptosis in Mammalian Aging. *Science* **309**, 481–484 (2005).

82. Depeint, F., Bruce, W. R., Shangari, N., Mehta, R. & O'Brien, P. J. Mitochondrial function and toxicity: Role of the B vitamin family on mitochondrial energy metabolism. *Chem. Biol. Interact.* **163**, 94–112 (2006).

83. Morscher, R. J. *et al.* Mitochondrial translation requires folate-dependent tRNA methylation. *Nature* **554**, 128–132 (2018).

84. Chou, Y.-F. & Huang, R.-F. S. Mitochondrial DNA deletions of blood lymphocytes as genetic markers of low folate-related mitochondrial genotoxicity in peripheral tissues. *Eur. J. Nutr.* **48**, 429–436 (2009).

85. Kronenberg, G. *et al.* Folate deficiency increases mtDNA and D-1 mtDNA deletion in aged brain of mice lacking uracil-DNA glycosylase. *Exp. Neurol.* **228**, 253–258 (2011).

86. Crider, K. S., Yang, T. P., Berry, R. J. & Bailey, L. B. Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role2. *Adv. Nutr.* **3**, 21–38 (2012).

87. Garcia, B. A., Luka, Z., Loukachevitch, L. V., Bhanu, N. V. & Wagner, C. Folate deficiency affects histone methylation. *Med. Hypotheses* **88**, 63–67 (2016).

88. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129**, 687–702 (1989).

89. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann. Epidemiol.* **1**, 263–276 (1991).

90. John, U. *et al.* Study of Health In Pomerania (SHIP): a health examination survey in an east German region: objectives and design. *Soz. Praventivmed.* **46**, 186–194 (2001).

91. Völzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).

92. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank | bioRxiv. Available at: https://www.biorxiv.org/content/10.1101/572347v1.supplementary-material. (Accessed: 14th August 2019)

93. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. *Nat. Genet.* **50**, 906–908 (2018).

94. Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

95. Herst, P. M., Rowe, M. R., Carson, G. M. & Berridge, M. V. Functional Mitochondria in Health and Disease. *Front. Endocrinol.* **8**, (2017).

96. Dai, D.-F., Rabinovitch, P. S. & Ungvari, Z. Mitochondria and cardiovascular aging. *Circ. Res.* **110**, 1109–1124 (2012).

97. Cui, H., Kong, Y. & Zhang, H. Oxidative stress, mitochondrial dysfunction, and aging. *J. Signal Transduct.* **2012**, 646354 (2012).

98. Sondheimer, N. *et al.* Neutral mitochondrial heteroplasmy and the influence of aging. *Hum. Mol. Genet.* **20**, 1653–1659 (2011).

99. Zhang, R., Wang, Y., Ye, K., Picard, M. & Gu, Z. Independent impacts of aging on mitochondrial DNA quantity and quality in humans. *BMC Genomics* **18**, 890 (2017).

100. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).

101. Lin, M. T. & Beal, M. F. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* **443**, 787 (2006).

102. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. *Circulation* **137**, e67–e492 (2018).

103. Ballinger Scott W. *et al.* Mitochondrial Integrity and Function in Atherogenesis. *Circulation* **106**, 544–549 (2002).

104. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* (2019). doi:10.1101/563866

105. Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data. *Trends Genet. TIG* **29**, 593–599 (2013).

106. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

107. Balboa, E. *et al.* MLN64 induces mitochondrial dysfunction associated with increased mitochondrial cholesterol content. *Redox Biol.* **12**, 274–284 (2017).

108. Bournat, J. C. & Brown, C. W. Mitochondrial Dysfunction in Obesity. *Curr. Opin. Endocrinol. Diabetes Obes.* **17**, 446–452 (2010).

109. Kimchi-Sarfaty, C. *et al.* A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525–528 (2007).

110. Goymer, P. Synonymous mutations break their silence. *Nat. Rev. Genet.* **8**, 92–92 (2007).

111. Stoekenbroek, R. M. *et al.* Heterogeneous impact of classic atherosclerotic risk factors on different arterial territories: the EPIC-Norfolk prospective population study. *Eur. Heart J.* **37**, 880–889 (2016).

CURRICULUM VITAE

Ryan Longchamps
rlongch1@jhmi.edu | (410) 440-8043
733 N Broadway, MRB 420, Baltimore, MD 21205

EDUCATION

Johns Hopkins School of Medicine | Baltimore, MD 8/2013 – 9/2019
Ph.D. Human Genetics
Elucidating the role mitochondrial DNA quantity and quality in cardiovascular disease risk

University of Maryland | College Park, MD 8/2007 – 5/2011
B.S. Physiology and Neurobiology

RESEARCH EXPERIENCE

Graduate Student | Johns Hopkins | Dan Arking 3/2014 – 9/2019

- Developed novel methods for calling mitochondrial DNA copy number from array data
- Evaluated factors, such as DNA extraction method, which affect mitochondrial DNA copy number estimation accuracy in the context of the current gold standards
- Analyzed and directed CHARGE mitochondrial DNA copy number GWAS efforts across 16 cohorts and 87,000 individuals
- Evaluated the role of mitochondrial heteroplasmy with overall mortality and cardiovascular disease in over 20,000 individuals from the CHARGE consortium
- Optimized methods for high throughput mitochondrial DNA sequencing enabling 384x multiplexing within a single flow-cell
- Directed lab efforts to develop allele-specific CRISPR-Cas9 edited cell lines targeting genes affecting mitochondrial DNA replication efficiency and fidelity

IRTA Postbac Fellow | NIH | Wendy Henderson 2/2012 – 5/2013

- Explored the use of microRNAs measured from whole blood as minimally invasive biomarkers for non-alcoholic fatty liver disease
- Optimized Ion Torrent PGM sequencing protocol for microbiome sequencing analyses
- Processed patient samples for future genetic, molecular and biochemical analyses
- Managed data repositories containing comprehensive patient information as well as laboratory budget and allotment of available spending

Research Assistant | U. of Maryland | Marco Colombini 8/2010 – 5/2011

- Investigated the role of ceramide channels in mitochondrial-mediated apoptosis through electrophysiological recordings

PUBLICATIONS

Sun J, **Longchamps RJ**, Piggott DA, Castellani CA, Sumpter JA, Brown TT, Mehta SH, Arking DE, Kirk GD. The association between HIV infection and mitochondrial DNA copy number in peripheral blood: a population-based, prospective cohort study. *J Infect Dis.* 2018 Nov 24. doi: 10.1093/infdis/jiy658.

Zhang Y, Guallar E, Ashar FN, **Longchamps RJ**, Castellani CA, Lane JA, Grove ML, Coresh J, Sotoodehnia N, Ilkhanoff L, Boerwinkle E, Pankratz N, Arking DE. Association of Mitochondrial DNA Copy Number and Sudden Cardiac Death: Findings from the Atherosclerosis Risk in Communities Study (ARIC). *Eur Heart J*. 2017 Dec 7;38(46):3443-3448. doi: 10.1093/eurheartj/ehx354.

Ashar FN, Zhang Y, **Longchamps RJ**, Lane JA, Moes A, Grove ML, Mychaleckyj JC, Taylor KD, Coresh J, Rotter JI, Boerwinkle E, Pankratz N, Guallar E, Arking DE. Mitochondrial DNA Copy Number is a Predictor of Cardiovascular Disease. *JAMA Cardiol.* 2017 Nov 1;2(11):1247-1255. doi: 10.1001/jamacardio.2017.3683.

MANUSCRIPTS IN PREPARATION

Longchamps RJ, Puiu D, Hong YS, Newcomb CE, Sumpter JA, Castellani CA, Grove ML, Walston JD, Windham BG, Coresh J, Boerwinkle E, Salzberg SL, Guallar E, Arking DE. Mitochondrial heteroplasmy is associated with overall mortality: findings from the Atherosclerosis Risk in Communities Study (ARIC).

Longchamps RJ, Lane JA, Grove ML, Lawson KS, Castellani CA, CHARGE aging and longevity working group, Boerwinkle E, Pankratz N, Arking DE. Genome-Wide Association Study of mitochondrial DNA copy number: the Cohorts for Heart & Aging Research in Genetic Epidemiology (CHARGE).

Longchamps RJ, Castellani CA, Newcomb CE, Sumpter JA, Grove ML, Guallar E, Pankratz N, Taylor KD, Rotter JI, Boerwinkle E, Arking DE. Evaluation of mitochondrial DNA copy number estimation techniques.

PRESENTATIONS

Longchamps RJ, Puiu D, Hong YS, Newcomb CE, Sumpter JA, Castellani CA, Grove ML, Walston JD, Windham BG, Coresh J, Boerwinkle E, Salzberg SL, Guallar E, Arking DE. (2018). Mitochondrial DNA Heteroplasmy is Associated with Overall Mortality. American Society for Human Genetics, 2018. Platform Presentation.

Longchamps, RJ, Mitchell, RN, Grove, ML, Boerwinkle, E, Arking, DE (2017). Explaining the Relationship Between mtDNA Quantity, Quality and Human Disease. Genetics Research Day, 2017. Poster. *2nd Place.*

Longchamps, RJ, Ashar, FN, Lane, JA, Bartz, TM, Pankratz, N, Arking, DE (2016). Comparison of Mitochondrial DNA Copy Number Estimation Techniques from Multiple Platforms. American Society for Human Genetics, 2016. Poster.

Longchamps, RJ, Coresh, J, Arking, DE Genome-Wide Interrogation of Spouse Selection Indicates Lack of Assortative Mating. Genetics Research Day, 2016. Poster. *Honorable Mention.*

GRANTS

Wolfe Street Competition ($10,000) 5/2017
A cohort study of mitochondrial DNA heteroplasmy and cardiovascular disease

HONORS

Leena Peltonen School of Human Genomics Trainee 8/2017
Wellcome Genome Campus
Maryland Genetics, Epidemiology, and Medicine Fellow 8/2015 – 9/2019
Johns Hopkins University
Intramural Research Training Award 2/2012 – 5/2013
National Institutes of Health

TEACHING

Graduate Student Tutor | Linkage and Association Analyses 3/2016 – 5/2018
Teaching Assistant | Computational Biology 11/2016 – 1/2018
Teaching Assistant | Advanced Topics in Human Genetics 11/2015 – 3/2016

SKILLS

R (excellent)
Bash (competent)
Plink
GWAS
Mendelian Randomization
Imputation
Data Analysis
Sequencing Analysis
qPCR
CRISPR-cas9 editing
Molecular Cloning
Cell Culture
Western Blot
DNA/RNA/Protein Extractions
Cluster/Cloud Computing